Topics in mathematical and applied statistics

Pas, S.L. van der

Topics in Mathematical and Applied Statistics

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 28 februari 2017
klokke 11.15 uur

door

Stéphanie Louise van der Pas
geboren te Hilversum in 1989

Promotor:	prof. dr. A.W. van der Vaart	Universiteit Leiden	
Promotiecommissie:	prof. dr. I. Castillo	Université Paris VI	
	prof. dr. R.D. Gill	Universiteit Leiden	
	prof. dr. P.D. Grünwald	Universiteit Leiden	(secretaris)
	dr. S. Gugushvili	Universiteit Leiden	
	dr. V. Ročková	University of Chicago	

Research supported by the Netherlands Organization for Scientific Research (NWO).

The Dutch Arthroplasty Register (Landelijke Registratie Orthopedische Implantaten / LROI) is gratefully acknowledged for providing the data analyzed in Chapter 6.

Contents

List of papers

The first five chapters of this thesis consist of the papers listed below, with minor changes to the references.

Chapter 1:
van der Pas, S.L., Kleijn, B.J.K. and van der Vaart, A.W. (2014), The horseshoe estimator: posterior concentration around nearly black vectors, *Electronic Journal of Statistics* **8**, 2585–2618.

Chapter 2:
van der Pas, S.L., Salomond, J.-B. and Schmidt-Hieber, J. (2016), Conditions for posterior contraction in the sparse normal means problem, *Electronic Journal of Statistics* **10**, 976–1000.

Chapter 3:
van der Pas, S., Szabó B. and van der Vaart, A., How many needles in the haystack? Adaptive inference and uncertainty quantification for the horseshoe. *Submitted.*

Chapter 4:
van der Pas, S.L. and van der Vaart, A.W., Bayesian community detection. *Submitted.*

Chapter 5:
van der Pas, S. and Grünwald, P., Almost the best of three worlds: risk, consistency and optional stopping for the switch criterion in nested model selection. To appear in *Statistica Sinica.*

The sixth chapter is based on material from the following two unpublished papers.

Chapter 6:
van der Pas, S.L., Nelissen, R.G.H.H. and Fiocco, M., Staged bilateral total joint arthroplasty patients in registries. Immortal time bias and methodological options. *Submitted.*

van der Pas, S.L., Nelissen, R.G.H.H., Schreurs, B.W. and Fiocco, M., Risk factors for early revision after unilateral and staged bilateral total hip replacement in the Dutch Arthroplasty Register. *In preparation.*

Introduction

This thesis is composed of papers on four topics: Bayesian theory for the sparse normal means problem (Chapters 1-3), Bayesian theory for community detection (Chapter 4), nested model selection (Chapter 5), and the application of competing risk methods in the presence of time-dependent clustering (Chapter 6). Each topic is briefly introduced in this Introduction.

Sparsity and shrinkage priors (Ch. 1 - 3)

A problem is sparse when there are only a few signals amidst a lot of noise. Those signals are like the proverbial needles in a haystack. The field of astronomy contributes many examples, such as supernovae detection (Clements et al., 2012). Other examples include the detection of genes associated to a certain disease (Silver et al., 2012) and image compression (Lewis and Knowles, 1992).

The particular sparse problem studied in the first three chapters of this thesis is the sparse normal means problem, also known as the sequence model. In the sparse normal means problem, a vector $Y^n \in \mathbb{R}^n$, $Y^n = (Y_1, Y_2, \ldots, Y_n)$, is observed, and assumed to have been generated according to the following model:

$$Y_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are assumed to be i.i.d. normally distributed with mean zero and known variance σ^2, and the vector of means $\theta \in \mathbb{R}^n$ is the parameter of interest. The sparsity assumption takes the form of assuming that θ is *nearly black*, meaning that almost all of its entries are zero. The number of nonzero entries in θ is denoted by p_n, a number which is assumed to increase with n, but not as fast as n: $p_n \to \infty, p_n = o(n)$. Other sparsity assumptions are possible, such as assuming that θ is in a strong or weak ℓ_s-ball for $s \in (0, 2)$ (Castillo and Van der Vaart, 2012; Johnstone and Silverman, 2004), but we do not pursue these further here.

The inferential goal can take several forms. *Recovery* of the parameter θ is one possible goal, and this is the main focus of Chapters 1 and 2. *Uncertainty quantification* is a second, and this is the topic of Chapter 3. A third goal, *model selection*, is not explored in this thesis, although the results in Chapter 3 do provide some avenues for further research.

There are many ways to achieve the aforementioned goals. The contributions of this thesis are in the field of frequentist Bayesian theory. The parameter of interest is equipped

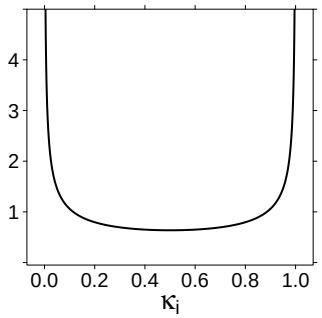

Figure 1: Prior density of κ_i for $\tau = 1$.

with a prior, which, when combined with the likelihood, leads to a posterior distribution, aspects of which we use to achieve our goals. We study the properties of the posterior from a frequentist point of view, meaning that we assume that there is some underlying true parameter that is generating the data.

The priors proposed for the sparse normal means problem are in general shrinkage priors, designed to yield many estimates close to or exactly equal to zero. The particular shrinkage prior studied in this thesis is the *horseshoe prior* (Carvalho et al., 2010). It has become popular, due to its good behaviour in simulation studies, and favorable theoretical properties (e.g. Armagan et al. (2013); Bhattacharya et al. (2014); Carvalho et al. (2010); Polson and Scott (2012*a*)). It has intuitive appeal, which can be explained through the origin of its name. The horseshoe prior is given by

$$\theta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2), \quad \lambda_i \sim C^+(0,1),$$

for $i = 1, \ldots, n$, where $C^+(0,1)$ is the standard half-Cauchy distribution. The parameter τ is a global parameter, shared by all means, while the parameter λ_i is a local parameter. How τ should be set is one of the main topics of Chapters 1 and 3. Regarding the name, if τ is known, we have the equality:

$$\mathbb{E}[\theta_i \mid Y_i = y_i, \tau] = (1 - \mathbb{E}[\kappa_i \mid Y_i = y_i, \tau])\, y_i,$$

where $\kappa_i = (1+\tau^2\lambda_i^2)^{-1}$, and $\mathbb{E}[\kappa_i \mid Y_i = y_i]$ can be interpreted as the amount of shrinkage towards zero. A half-Cauchy prior on λ_i implies a $\text{Be}(\frac{1}{2}, \frac{1}{2})$ prior on κ_i in the special case when $\tau = 1$. The horseshoe prior is named after the $\text{Be}(\frac{1}{2}, \frac{1}{2})$ prior, which resembles a horseshoe (Figure 1).

The intuitive appeal lies in the concentration of mass near zero and one, which caters to the true signals and the nonzero means respectively. Decreasing τ leads to more prior mass near one, corresponding to more shrinkage. One of the main contributions of this thesis is the guideline that optimal recovery (in the minimax sense) can be achieved by setting τ at most of the order $(p_n/n)\sqrt{\log(n/p_n)}$ (Chapter 1).

In practice, the number p_n is unknown and thus there is a need for a procedure that adapts to the unknown sparsity level. In Chapter 3, two ways of handling τ are considered.

The first is empirical Bayes, where τ is estimated based on the data and the resulting value plugged into the prior. The second is hierarchical Bayes, where τ receives a hyperprior. In both cases, recovery of θ is possible at the near-minimax rate, and credible balls are both honest (they contain the true value with some prescribed probability) and adaptive (they are as small as possible), under some conditions. In addition, credible intervals are guaranteed to have good coverage if the underlying signal is either 'small' or 'large', but almost surely do not contain the truth if the signal is close to approximately $\sqrt{2 \log n}$.

The results of Chapter 1 led to the question what properties of the horseshoe prior make it so well suited for recovery, and whether it is unique in that regard. The results from Chapter 2 show that the horseshoe is not that special: many priors in the class of scale mixtures of normals enjoy the same good behaviour. We provide conditions under which the posterior contracts at the minimax rate. Recovery of the nonzeroes requires tails that are at least exponential; recovery of the zeroes requires sufficient mass close to zero, and not too much mass in the interval $[(p_n/n) \log(n/p_n), 1]$. Many priors satisfy these conditions. However, the horseshoe may be special after all, because it represents a boundary case with respect to the thickness of its tails. This could explain the good coverage properties of the horseshoe's credible balls. Whether the uncertainty quantification properties of the horseshoe, as described in Chapter 3, can be generalized to scale mixtures of normals is an open question.

An attractive property of the horseshoe is that some of the aspects of its posterior can be easily and quickly computed, without the need for MCMC. Functions for the horseshoe's posterior mean, posterior variance, the MMLE and credible intervals are available in the R package 'horseshoe' (van der Pas et al., 2016).

Community detection (Ch. 4)

In this chapter, like the previous ones, Bayesian posterior distributions are studied from a frequentist point of view, but unlike the previous chapters, the theory is for data with a network structure. The aim is to detect communities in, for example, a social network. The network is assumed to be generated according to the *stochastic block model*, in which the probability of the existence of a connection between two individuals (nodes) only depends on each individual's community membership.

We equip all parameters of the stochastic block models with priors, and use the posterior mode as an estimator of the community memberships (MAP-estimation). We call the resulting estimator the *Bayesian modularity*, following Bickel et al. (2009). Two instances are studied. In the first, the *dense* situation, the probabilities of connections between individuals remain fixed. The second and most complicated situation is the *sparse* situation, in which the probability of a connection between two individuals tends to zero as the network grows in size.

Weak and strong consistency are proven, the former meaning that only a fraction of the nodes are misclassified, and the latter that none of the nodes are misclassified. The theorems require the assumption that the expected degree is at least of order $\log^2 n$, where n is the number of nodes in the network. Whether this assumption can be weakened to an expected degree of order $\log n$ remains an open question.

Nested model selection (Ch. 5)

In Chapter 5, we turn to model selection. The models under consideration are nested exponential family models. For example, one model could consist of all univariate normal distributions with unknown mean and unknown variance, while the other model only contains the standard normal distribution.

Optimality of a model selection criterion can be defined in many different ways. Three of them are discussed in this Chapter: consistency, minimax rate optimality, and robustness to optional stopping. The switch criterion, a new model selection criterion based on the switch distribution introduced by Van Erven et al. (2012), is evaluated on those three properties.

Consistency guarantees that if the data is actually generated according to one of the models, then that model will be selected eventually. *minimax rate optimality* is a measure of the accuracy of the parameter estimation step that follows the model selection. minimax rate optimality and consistency are mutually exclusive properties (Yang, 2005). The main contribution of Chapter 5 is that the switch criterion is consistent while missing the minimax risk by a factor of order $\log \log n$, if the criterion is used in combination with efficient estimators of the parameters.

The third property, robustness to optional stopping, has attracted attention because most standard null hypothesis significance tests which output a p-value, do not have this property (Armitage et al., 1969; Wagenmakers, 2007). In the classical framework, a researcher has to decide the sample size in advance. This guideline is not always adhered to; in a recent survey of psychologists, approximately 55% of participants admitted to deciding whether to collect more data after looking at their results to see if they were significant (John et al., 2012). If a criterion is robust to optional stopping, the validity of the results will not be affected by the use of such stopping rules. As discussed in Chapter 5, the switch criterion is robust to optional stopping, if the null hypothesis is a point hypothesis. Thus, the switch criterion comes close to achieving all three desirable properties.

Hip arthroplasty data and bilateral patients (Ch. 6)

The final chapter of this thesis is on the topic of hip arthroplasty registry data, and is of a different character than the preceding ones. It is the result of an ongoing collaboration with Marta Fiocco, Rob Nelissen and Wim Schreurs. The goal is to determine which patient characteristics are associated with time to revision surgery after hip replacement surgery, using the data collected by the LROI (Landelijke Registratie Orthopedische Implantaten / Dutch Arthroplasty Register).

Total hip arthroplasty (THA) is a common procedure in The Netherlands; the LROI registers approximately 28.000 THAs annually, in most cases following an ostheoarthritis diagnosis (LROI, 2014). After the primary surgery, there may be a need for revision surgery, which is defined as any change (insertion, replacement, and/or removal) of one or more components of a prosthesis. Revision may be required due to several reasons, such as mechanical loosening, infection and fracture.

Many patients will have not one, but both hip joints replaced and thus receive bilateral

prostheses during the postoperative course of their first hip or knee arthroplasty. In 2014, 20% of THAs in The Netherlands concerned the placement of a second prosthesis (LROI, 2014). Bilateral patients have been theorized to have different risks of revision compared to unilateral patients, as the two hips may affect each other regarding loosening (Buchholz et al., 1985). In addition, although the primary diagnosis for surgery may be osteoarthritis, patients with several total joint arthroplasties within a short time period may reflect a different patient population compared to a patient who has only one implant during a, say, five year follow-up.

There are some methodological difficulties in studying bilateral patients. First of all, the observations contributed by a bilateral patient are dependent. A second complication is that a patient may become bilateral at any point in time after the first surgery. This makes subgroup analysis problematic, as there is a risk of immortal time bias (e.g. Oscar winners 'live longer' because one needs to be alive to win an Oscar (Sylvestre et al., 2006)).

In addition, any patient may die before experiencing revision of the implant. If this competing risk of death is not appropriately accounted for, the risk of revision surgery will be overestimated (Keurentjes et al., 2012; Ranstam et al., 2011). This is especially important for these analyses given the age of most patients: the average age at index surgery is 69 years for THA (LROI, 2014). In Chapter 6 the aforementioned complications are explained in detail, and methods that have been proposed to handle them are reviewed. The chapter is concluded with some preliminary analyses of the LROI data.

1

Posterior concentration of the horseshoe around nearly black vectors

Abstract

We consider the horseshoe estimator due to Carvalho et al. (2010) for the multivariate normal mean model in the situation that the mean vector is sparse in the nearly black sense. We assume the frequentist framework where the data is generated according to a fixed mean vector. We show that if the number of nonzero parameters of the mean vector is known, the horseshoe estimator attains the minimax ℓ_2 risk, possibly up to a multiplicative constant. We provide conditions under which the horseshoe estimator combined with an empirical Bayes estimate of the number of nonzero means still yields the minimax risk. We furthermore prove an upper bound on the rate of contraction of the posterior distribution around the horseshoe estimator, and a lower bound on the posterior variance. These bounds indicate that the posterior distribution of the horseshoe prior may be more informative than that of other one-component priors, including the Lasso.

This chapter has appeared as S.L. van der Pas, B.J.K. Kleijn and A.W. van der Vaart (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

1.1 Introduction

We consider the normal means problem, where we observe a vector $Y \in \mathbb{R}^n$, $Y = (Y_1, \ldots, Y_n)$, such that

$$Y_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

for independent normal random variables ε_i with mean zero and variance σ^2. The vector $\theta = (\theta_1, \ldots, \theta_n)$ is assumed to be sparse, in the 'nearly black' sense that the number of nonzero means

$$p_n := \#\{i : \theta_i \neq 0\}$$

is $o(n)$ as $n \to \infty$. A natural Bayesian approach to recovering θ would be to induce sparsity through a 'spike and slab' prior (Mitchell and Beauchamp, 1988), which consists of a mixture of a Dirac measure at zero and a (heavy-tailed) continuous distribution. Johnstone and Silverman (2004) analyzed an empirical Bayes version of this approach, where the mixing weight is obtained by marginal maximum likelihood. In the frequentist setup that the data are generated according to a fixed mean vector, they showed that the empirical Bayes coordinatewise posterior median attains the minimax rate, in ℓ_q norm, $q \in (0, 2]$, for mean vectors that are either nearly black or of bounded ℓ_p norm, $p \in (0, 2]$. Castillo and Van der Vaart (2012) analyzed a fully Bayesian version, where the proportion of nonzero coefficients is modelled by a prior distribution. They identified combinations of priors on this proportion and on the nonzero coefficients (the 'slab') that yield posterior distributions concentrating around the underlying mean vector at the minimax rate in ℓ_q norm, $q \in (0, 2]$, for mean vectors that are nearly black, and in ℓ_q norm, $q \in (0, 2)$ for mean vectors of bounded weak ℓ_p norm, $p \in (0, q)$. Other work on empirical Bayes approaches to the two-group model includes (Efron, 2008; Jiang and Zhang, 2009; Yuan and Lin, 2005).

As a full Bayesian approach with a mixture of a Dirac and a continuous component may require exploration of a model space of size 2^n, implementation on large datasets is currently impractical, although Castillo and Van der Vaart (2012) present an algorithm which can compute several aspects of the posterior in polynomial time, provided sufficient memory can be allocated. Several authors, including (Armagan et al., 2013; Griffin and Brown, 2010), have proposed one-component priors, which model the spike at zero by a peak in the prior density at this point. For most of these proposals, theoretical justification in terms of minimax risk rates or posterior contraction rates is lacking. The Lasso estimator (Tibshirani, 1996), which arises as the MAP estimator after placing a Laplace prior with common parameter on each θ_i, is an exception. It attains close to the minimax risk rate in ℓ_q, $q \in [1, 2]$ (Bickel et al. (2009)). It has however been recently shown that the corresponding full posterior distribution contracts at a much slower rate than the mode (Castillo et al., 2015). This is undesirable, because this implies that the posterior distribution cannot provide an adequate measure of uncertainty in the estimate.

In general one would use a posterior distribution both for recovery and for uncertainty quantification. For the first, a measure of centre, such as a median or mode, suffices. For the second, one typically employs a credible set, which is defined as a central set of prescribed posterior probability. For realistic uncertainty quantification it is necessary that the posterior contracts to its center at the same rate as the posterior median or mode approaches the true parameter.

In this paper we study the posterior distribution resulting from the horseshoe prior, which is a one-component prior, introduced in (Carvalho et al., 2009, 2010) and expanded upon in (Polson and Scott, 2012*a*,*b*; Scott, 2011). It combines a pole at zero with Cauchy-like tails. The corresponding estimator does not face the computational issues of the point mass mixture models. Carvalho et al. (2010) already showed good behaviour of the horseshoe estimator in terms of Kullback-Leibler risk when the true mean is zero. Datta and Ghosh (2013) proved some optimality properties of a multiple testing rule induced by the horseshoe estimator. In this paper, we prove that the horseshoe estimator achieves the minimax quadratic risk, possibly up to a multiplicative constant. We furthermore prove that the posterior variance is of the order of the minimax risk, and thus the posterior contracts at the minimax rate around the underlying mean vector. These results are proven under the assumption that the number p_n of nonzero parameters is known. However, we also provide conditions under which the horseshoe estimator combined with an empirical Bayes estimator still attains the minimax rate, when p_n is unknown.

This paper is organized as follows. In Section 1.2, the horseshoe prior is described and a summary of simulation results is given. The main results, that the horseshoe estimator attains the minimax squared error risk (up to a multiplicative constant) and that the posterior distribution contracts around the truth at the minimax rate, are stated in Section 1.3. Conditions on an empirical Bayes estimator of the key parameter τ such that the minimax ℓ_2 risk will still be obtained are given in Section 1.4. The behaviour of such an empirical Bayes estimate is compared to a full Bayesian version in a numerical study in Section 1.5. Section 1.6 contains some concluding remarks. The proofs of the main results and supporting lemmas are in the appendix.

1.1.1 Notation

We write $A_n \asymp B_n$ to denote $0 < \lim_{n\to\infty} \inf \frac{A_n}{B_n} \leq \lim_{n\to\infty} \sup \frac{A_n}{B_n} < \infty$ and $A_n \lesssim B_n$ to denote that there exists a positive constant c independent of n such that $A_n \leq cB_n$. $A \vee B$ is the maximum of A and B, and $A \wedge B$ the minimum of A and B. The standard normal density and cumulative distribution are denoted by ϕ and Φ and we set $\Phi^c = 1 - \Phi$. The norm $\|\cdot\|$ will be the ℓ_2 norm and the class of nearly black vectors will be denoted by $\ell_0[p_n] := \{\theta \in \mathbb{R}^n : \#(1 \leq i \leq n : \theta_i \neq 0) \leq p_n\}$.

1.2 The horseshoe prior

In this section, we give an overview of some known properties of the horseshoe estimator which will be relevant to the remainder of our discussion. The horseshoe prior for a parameter θ modelling an observation $Y \sim \mathcal{N}(\theta, \sigma^2 I_n)$ is defined hierarchically (Carvalho et al., 2010):

$$\theta_i \mid \lambda_i, \tau \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2), \quad \lambda_i \sim C^+(0,1),$$

for $i = 1, \ldots, n$, where $C^+(0,1)$ is a standard half-Cauchy distribution. The parameter τ is assumed to be fixed in this paper, rendering the θ_i independent *a priori*. The corresponding density p_τ increases logarithmically around zero, while its tails decay quadratically. The

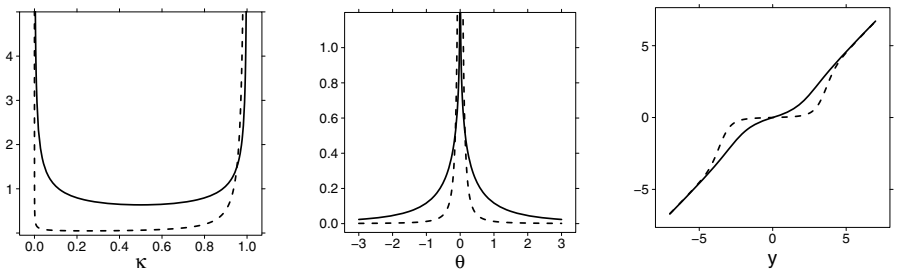

Figure 1.1: The effect of decreasing τ on the priors on κ (left) and θ (middle) and the posterior mean $T_\tau(y)$ (right). The solid line corresponds to $\tau = 1$, the dashed line to $\tau = 0.05$. Decreasing τ results in a higher prior probability of shrinking the observations towards zero.

posterior density of θ_i given λ_i and τ is normal with mean $(1-\kappa_i)y_i$, where $\kappa_i = \frac{1}{1+\tau^2\lambda_i^2}$. Hence, by Fubini's theorem:

$$\mathbb{E}[\theta_i \mid y_i, \tau] = (1 - \mathbb{E}[\kappa_i \mid y_i, \tau])y_i.$$

The posterior mean $\mathbb{E}[\theta \mid y, \tau]$ will be referred to as the horseshoe estimator and denoted by $T_\tau(y)$. The horseshoe prior takes its name from the prior on κ_i, which is given by:

$$p_\tau(\kappa_i) = \frac{\tau}{\pi} \frac{1}{1-(1-\tau^2)\kappa_i} (1-\kappa_i)^{-\frac{1}{2}} \kappa_i^{-\frac{1}{2}}.$$

If $\tau = 1$, this reduces to a $\text{Be}(\frac{1}{2}, \frac{1}{2})$ distribution, which looks like a horseshoe. As illustrated in Figure 1.1, decreasing τ skews the prior distribution on κ_i towards one, corresponding to more mass near zero in the prior on θ_i and a stronger shrinkage effect in $T_\tau(y)$.

The posterior mean can be expressed as:

$$T_\tau(y_i) = y_i \left(1 - \frac{2\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}; \frac{y_i^2}{2\sigma^2}, 1-\frac{1}{\tau^2}\right)}{3\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}; \frac{y_i^2}{2\sigma^2}, 1-\frac{1}{\tau^2}\right)}\right) = y_i \frac{\int_0^1 z^{\frac{1}{2}} \frac{1}{\tau^2+(1-\tau^2)z} e^{\frac{y_i^2}{2\sigma^2}z} dz}{\int_0^1 z^{-\frac{1}{2}} \frac{1}{\tau^2+(1-\tau^2)z} e^{\frac{y_i^2}{2\sigma^2}z} dz}, \tag{1.1}$$

where $\Phi_1(\alpha, \beta, \gamma; x, y)$ denotes the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965).

An unanswered question so far has been how τ should be chosen. Intuitively, τ should be small if the mean vector is very sparse, as the horseshoe prior will then place more of its mass near zero. By approximating the posterior distribution of τ^2 given $\kappa = (\kappa_1, \ldots, \kappa_n)$ in case a prior on τ is used, Carvalho et al. (2010) show that if most observations are shrunk near zero, τ will be very small with high probability. They suggest a half-Cauchy prior on τ. Datta and Ghosh (2013) implemented this prior on τ and their plots of posterior draws for τ at various sparsity levels indicate the expected relationship between τ and the sparsity level: the posterior distribution of τ tends to concentrate around smaller values when the underlying mean vector is sparser. As will be discussed further in the next

section, the value $\tau = \frac{p_n}{n}$ (up to a log factor) is optimal in terms of mean square error and posterior contraction rates.

In case τ is estimated empirically, as will be considered in Section 1.4, the horseshoe estimator can be computed by plugging this estimate into expression (1.1), thereby avoiding the use of MCMC. Other aspects of the posterior, such as the posterior variance, can be computed using such a plug-in procedure as well. Polson and Scott (2012*a*) and Polson and Scott (2012*b*) consider computation of the horseshoe estimator based on the representation in terms of degenerate hypergeometric functions, as these can be efficiently computed using converging series of confluent hypergeometric functions. They report unproblematic computations for τ^2 between $\frac{1}{1000}$ and 1000. A second option is to apply a quadrature routine to the integral representation in (1.1). As the continuity and symmetry of $T_\tau(y)$ in y can be taken advantage of when computing the horseshoe estimator for a large number of observations, the complexity of these computations mostly depends on the value of τ. Both approaches will be slower for smaller values of τ. Hence, if we use the (estimated) sparsity level $\frac{p_n}{n}$ (up to a log factor) for τ, the computation of the horseshoe estimator will be slower if there are fewer nonzero parameters. As noted by Scott (2010), problems arise in Gibbs sampling precisely when τ is small as well. Hence care needs to be taken with any computational approach if $\frac{p_n}{n}$ is suspected to be very close to zero.

The performance of the horseshoe prior, with additional priors on τ and σ^2, in various simulation studies has been very promising. Carvalho et al. (2010) simulated sparse data where the nonzero components were drawn from a Student-t density and found that the horseshoe estimator systematically beat the MLE, the double-exponential (DE) and normal-exponential-gamma (NEG) priors, and the empirical Bayes model due to Johnstone and Silverman (2004) in terms of square error loss. Only when the signal was neither sparse nor heavy-tailed did the MLE, DE and NEG priors have an edge over the horseshoe estimator. In similar experiments in (Carvalho et al., 2009; Polson and Scott, 2012*a*) the horseshoe prior outperformed the DE prior, while behaving similarly to a heavy-tailed discrete mixture. In a wavelet-denoising experiment under several noise levels and loss functions, the horseshoe estimator compared favorably to the discrete wavelet transform and the empirical Bayes model (Polson and Scott, 2010). Bhattacharya et al. (2012) applied several shrinkage priors to data with the underlying mean vector consisting of zeroes and fixed nonzero values and found the posterior median of the horseshoe prior performing better in terms of squared error than the Bayesian Lasso (BL), the Lasso, the posterior median of a point mass mixture prior as in (Castillo and Van der Vaart, 2012) and the empirical Bayes model proposed by Johnstone and Silverman (2004), and comparable to their proposed Dirichlet-Laplace (DL) prior with parameter $\frac{1}{n}$. Results in (Armagan et al., 2013) are similar. In a second simulation setting, Bhattacharya et al. (2012) generated data of length $n = 1000$, with the first ten means equal to 10, the next 90 equal to a number $A \in \{2, \ldots, 7\}$ and the remainder equal to zero. In this simulation, the horseshoe prior beat the BL (except when $A = 2$) and the DL prior with parameter $\frac{1}{n}$ (except when $A = 7$), while performing similarly to the DL prior with parameter $\frac{1}{2}$. It is worthy of note that Koenker (2014) generated data according to the same scheme and applied the empirical Bayes procedures due to Martin and Walker (2014) (EBMW) and Koenker and Mizera (2014) (EBKM) to it. The MSE of EBMW was lower than that of the horseshoe prior for $A \in \{5, 6, 7\}$, while that of EBKM was much lower in all cases.

1.3 Mean square error and bounds on the posterior variance

In this section, we study the mean square error of the horseshoe estimator, and the spread of the posterior distribution, under the assumption that the number of nonzero parameters p_n is known. Theorem 1.1 provides an upper bound on the mean square error, and shows that for a range of choices of the global parameter τ, the horseshoe estimator attains the minimax ℓ_2 risk, possibly up to a multiplicative constant. Theorem 1.3 states upper bounds on the rate of contraction of the posterior distribution around the underlying mean vector and around the horseshoe estimator, again for a range of values of τ. These upper bounds are equal, up to a multiplicative constant, to the minimax risk. The contraction rate around the truth is sharp, but this may not be the case for the rate of contraction around the horseshoe estimator. Theorems 1.4 and 1.5 provide more insight into the spread of the posterior distribution for various values of τ and indicate that $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ is a good choice.

Theorem 1.1. *Suppose $Y \sim \mathcal{N}(\theta_0, \sigma^2 I_n)$. Then the estimator $T_\tau(y)$ satisfies*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|T_\tau(Y) - \theta_0\|^2 \lesssim p_n \log \frac{1}{\tau} + (n - p_n)\tau\sqrt{\log \frac{1}{\tau}} \tag{1.2}$$

for $\tau \to 0$, as $n, p_n \to \infty$ and $p_n = o(n)$.

By the minimax risk result in (Donoho et al., 1992), we also have a lower bound:

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|T_\tau(Y) - \theta_0\|^2 \geq 2\sigma^2 p_n \log \frac{n}{p_n}(1 + o(1)),$$

as $n, p_n \to \infty$ and $p_n = o(n)$. The choice $\tau = (\frac{p_n}{n})^\alpha$, for $\alpha \geq 1$, leads to an upper bound (1.2) of order $p_n \log(n/p_n)$, with (as can be seen from the proof) a multiplicative constant of at most $4\alpha\sigma^2$. Thus, for this choice of τ, we have:

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|T_\tau(Y) - \theta_0\|^2 \asymp p_n \log \frac{n}{p_n}.$$

The horseshoe estimator therefore performs well as a point estimator, as it attains the minimax risk (possibly up to a multiplicative constant of at most 2 for $\alpha = 1$). This may seem surprising, as the prior does not include a point mass at zero to account for the assumed sparsity in the underlying mean vector. Theorem 1.1 shows that the pole at zero of the horseshoe prior mimics the point mass well enough, while the heavy tails ensure that large observations are not shrunk too much.

An upper bound on the rate of contraction of the posterior can be obtained through an upper bound on the posterior variance. The posterior variance can be expressed as:

$$\operatorname{var}(\theta_i \mid y_i) = \frac{\sigma^2}{y_i} T_\tau(y_i) - (T_\tau(y_i) - y_i)^2 + y_i^2 \frac{8\Phi_1\left(\frac{1}{2}, 1, \frac{7}{2}; \frac{y_i^2}{2\sigma^2}, 1 - \frac{1}{\tau^2}\right)}{15\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}; \frac{y_i^2}{2\sigma^2}, 1 - \frac{1}{\tau^2}\right)}.$$

Details on the computation can be found in Lemma 1.10. Using a similar approach as when bounding the ℓ_2 risk, we can find an upper bound on the expected value of the posterior variance.

Theorem 1.2. *Suppose $Y \sim \mathcal{N}(\theta_0, \sigma^2 I_n)$. Then the variance of the posterior distribution corresponding to the horseshoe prior satisfies*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \sum_{i=1}^{n} \operatorname{var}(\theta_{0i} \mid Y_i) \lesssim p_n \log \frac{1}{\tau} + (n - p_n)\tau \sqrt{\log \frac{1}{\tau}} \tag{1.3}$$

for $\tau \to 0$, as $n, p_n \to \infty$ and $p_n = o(n)$.

Again, the choice $\tau = (\frac{p_n}{n})^\alpha$, for $\alpha \geq 1$ leads to an upper bound (1.3) of the order of the minimax risk. This result indicates that the posterior contracts fast enough to be able to provide a measure of uncertainty of adequate size around the point estimate. Theorems 1.1 and 1.2 combined allow us to find an upper bound on the rate of contraction of the full posterior distribution, both around the underlying mean vector and around the horseshoe estimator.

Theorem 1.3. *Under the assumptions of Theorem 1.1, with $\tau = (\frac{p_n}{n})^\alpha$, $\alpha \geq 1$:*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_\tau \left(\theta : \|\theta - \theta_0\|^2 > M_n p_n \log \frac{n}{p_n} \;\middle|\; Y \right) \to 0, \tag{1.4}$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_\tau \left(\theta : \|\theta - T_\tau(Y)\|^2 > M_n p_n \log \frac{n}{p_n} \;\middle|\; Y \right) \to 0, \tag{1.5}$$

for every $M_n \to \infty$ as $n \to \infty$.

Proof. Combine Markov's inequality with the results of Theorems 1.1 and 1.2 for (1.4), and only with the result of Theorem 1.2 for (1.5). □

A remarkable aspect of the preceding Theorems is that many choices of τ, such as $\tau = (\frac{p_n}{n})^\alpha$ for any $\alpha \geq 1$, lead to an upper bound of the order $p_n \log(n/p_n)$ on the worst case ℓ_2 risk and posterior contraction rate. The upper bound on the rate of contraction in (1.4) is sharp, as the posterior cannot contract faster than the minimax rate around the true mean vector (Ghosal et al., 2000). However, this is not necessarily the case for the upper bound in (1.5), and for $\tau = (\frac{p_n}{n})^\alpha$ with $\alpha > 1$, the posterior spread may be of smaller order than the rate at which the horseshoe estimator approaches the underlying mean vector. Theorems 1.4 and 1.5 provide more insight into the effect of choosing different values of τ on the posterior spread and mean square error.

Theorem 1.4. *Suppose $Y \sim \mathcal{N}(\theta_0, \sigma^2 I_n)$, $\theta_0 \in \ell_0[p_n]$. Then the variance of the posterior distribution corresponding to the horseshoe prior satisfies*

$$\inf_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \sum_{i=1}^{n} \operatorname{var}(\theta_{0i} \mid Y_i) \gtrsim (n - p_n)\tau \sqrt{\log \frac{1}{\tau}} \tag{1.6}$$

for $\tau \to 0$ and $p_n = o(n)$, as $n \to \infty$. This lower bound is sharp for vectors $\theta_{0,n}$ with p_n entries equal to a_n and the remaining entries equal to zero, if a_n is such that $|a_n| \lesssim 1/\sqrt{\log(1/\tau)}$.

Theorem 1.5. *Suppose $Y \sim \mathcal{N}(\theta_{0,n}, \sigma^2 I_n)$ and $\theta_{0,n} \in \ell_0[p_n]$ is such that p_n entries are equal to $\gamma\sqrt{2\sigma^2 \log(1/\tau)}$, $\gamma \in (0,1)$, and all remaining entries are equal to zero. Then:*

$$\mathbb{E}_{\theta_{0,n}} \|T_\tau(Y) - \theta_{0,n}\|^2 \asymp p_n \log \frac{1}{\tau} + (n - p_n)\tau\sqrt{\log \frac{1}{\tau}}, \tag{1.7}$$

and

$$\mathbb{E}_{\theta_{0,n}} \sum_{i=1}^{n} \operatorname{var}(\theta_{0,ni} \mid Y_i) \asymp p_n \tau^{(1-\gamma)^2} \left(\log \frac{1}{\tau}\right)^{\gamma - \frac{1}{2}} + (n - p_n)\tau\sqrt{\log \frac{1}{\tau}}, \tag{1.8}$$

for $\tau \to 0$ and $p_n = o(n)$, as $n \to \infty$.

Consider $\tau = (\frac{p_n}{n})^\alpha$. Three cases can be discerned:

(i) $0 < \alpha < 1$. Lower bound (1.6) may exceed the minimax rate, implying suboptimal spread of the posterior distribution in the squared ℓ_2 sense.

(ii) $\alpha = 1$. Bounds (1.3) and (1.6) differ by a factor $\sqrt{\log(n/p_n)}$, as do (1.7) and (1.8). The gap can be closed by choosing $\tau = \frac{p_n}{n}\sqrt{\log \frac{n}{p_n}}$.

(iii) $\alpha > 1$. Bound (1.6) is not very informative, but Theorem 1.5 exhibits a sequence $\theta_{0,n} \in \ell_0[p_n]$ for which there is a mismatch between the order of the mean square error and the posterior variance. Bounds (1.7) and (1.8) are of the orders $p_n(\log(1/\tau) + \tau^{1-1/\alpha}\sqrt{\log(1/\tau)})$ and $p_n(\tau^{(1-\gamma)^2}(\log(1/\tau))^{\gamma-1/2} + \tau^{1-1/\alpha}\sqrt{\log(1/\tau)})$, respectively. Hence up to logarithmic factors the total posterior variance (1.8) is a factor $\tau^{(1-1/\alpha)\wedge(1-\gamma)^2}$ smaller than the square distance of the center of the posterior to the truth (1.7). For $p_n \leq n^c$ for some $c > 0$, this factor behaves as a power of n.

These observations suggest that $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ is a good choice, because then (1.2), (1.3), (1.6), (1.7), (1.8) are all of the order $p_n \log(n/p_n)$, suggesting that the posterior contracts at the minimax rate around both the truth and the horseshoe estimator.

1.4 Empirical Bayes estimation of τ

A natural follow-up question is how to choose τ in practice, when p_n is unknown. As discussed in Section 1.2, the full Bayesian approach suggested by Carvalho et al. (2010) performs well in simulations. The analysis of such a hierarchical prior would however require different tools than the ones we have used so far. An empirical Bayes estimate of τ would be a natural solution, and allows us in practice to use one of the representations in (1.1) for computations, instead of an MCMC-type algorithm.

By adapting the approach in Paragraph 6.2 in (Johnstone and Silverman, 2004), we can find conditions under which the horseshoe estimator with an empirical Bayes estimate of τ will still attain the minimax ℓ_2 risk. Based on the consideration of Section 1.3, we proceed with the choices $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ and $\tau = \frac{p_n}{n}$. The former is optimal in the sense that the posterior spread is of the order of the minimax risk, but the latter has the simple interpretation of being the proportion of nonzero means, and the difference between the two is only the square root of a log factor.

Theorem 1.6. *Suppose we observe an n-dimensional vector $Y \sim \mathcal{N}(\theta_0, \sigma^2 I_n)$ and we use $T_{\widehat{\tau}}(y)$ as our estimator of θ_0. If $\widehat{\tau} \in (0,1)$ satisfies the following two conditions for $\tau = \frac{p_n}{n}$ or $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$:*

1. *$\mathbb{P}_{\theta_0}(\widehat{\tau} > c\tau) \lesssim \frac{p_n}{n}$ for a constant $c \geq 1$ such that $\tau \leq \frac{1}{c}$;*
2. *There exists a function $g : \mathbb{N} \times \mathbb{N} \to (0,1)$ such that $\widehat{\tau} \geq g(n, p_n)$ with probability one and $-\log(g(n,p_n))\mathbb{P}_{\theta_0}(\widehat{\tau} \leq \tau) \lesssim \log(n/p_n)$,*

then:

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \|T_{\widehat{\tau}}(Y) - \theta_0\|^2 \asymp p_n \log \frac{n}{p_n} \tag{1.9}$$

as $n, p_n \to \infty$ and $p_n = o(n)$. If only the first condition can be verified for an estimator $\widehat{\tau}$, then $\sup\{\frac{1}{n}, \widehat{\tau}\}$ will have an ℓ_2 risk of at most order $p_n \log n$.

The first condition requires that $\widehat{\tau}$ does not overestimate the fraction $\frac{p_n}{n}$ of nonzero means (up to a log factor) too much or with a too large probability. If $p_n \geq 1$, as we have assumed, then it is satisfied already by $\widehat{\tau} = \frac{1}{n}$ (and $c = 1$). According to the last assertion of the theorem, this 'universal threshold' yields the rate $p_n \log n$ (possibly up to a multiplicative constant). This is equal to the rate of the Lasso estimator with the usual choice of $\lambda = 2\sqrt{2\sigma^2 \log n}$ (Bickel et al., 2009). However, in the framework where $p_n \to \infty$, the estimator $\widehat{\tau} = \frac{1}{n}$ will certainly underestimate the sparsity level. A more natural estimator of $\frac{p_n}{n}$ is:

$$\widehat{\tau} = \frac{\#\{|y_i| \geq \sqrt{c_1 \sigma^2 \log n}, i = 1, \ldots, n\}}{c_2 n}, \tag{1.10}$$

where c_1 and c_2 are positive constants. By Lemma 1.13, this estimator satisfies the first condition for $\tau = \frac{p_n}{n}$ and $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ if $c_1 > 2, c_2 > 1$ and $p_n \to \infty$ or $c_1 = 2, c_2 > 1$ and $p_n \gtrsim \log n$. Thus $\max\{\widehat{\tau}, \frac{1}{n}\}$ will also lead to a rate of at most order $p_n \log n$ under these conditions. Its behaviour will be explored further in Section 1.5.

The rate can be improved to $p_n \log(n/p_n)$ if the second condition is met as well, which ensures that the sparsity level is not underestimated too much or by a too large probability. As we are not aware of any estimators meeting this condition for all θ_0, this condition is currently mostly of theoretical interest. If the true mean vector is very sparse, in the sense that there are relatively few nonzero means or the nonzero means are close to zero, there is not much to be gained in terms of rates by meeting this condition. The extra occurrence of p_n relative to the rate $p_n \log n$ is of interest only if p_n is relatively large. For instance, if $p_n \asymp n^\alpha$ for $\alpha \in (0,1)$, then $p_n \log(n/p_n) = (1-\alpha) p_n \log n$, which suggests a decrease of the proportionality constant in (1.9), particularly if α is close to one. Furthermore, when p_n is large, the constant in (1.9) may be sensitive to the fine properties of $\widehat{\tau}$, as it depends on $g(n, p_n)$ (as can be seen in the proof). If $\widehat{\tau}$ seriously underestimates the sparsity level, the corresponding value of $g(n, p_n)$ from the second condition may be so small that the upper bound on the multiplicative constant before (1.9) becomes very large. Hence in this case, $\widehat{\tau}$ is required to be close to the proportion $\frac{p_n}{n}$ (up to a log factor) with large probability in order to get an optimal rate.

Datta and Ghosh (2013) warned against the use of an empirical Bayes estimate of τ for the horseshoe prior, because the estimate might collapse to zero. Their references for this statement, Scott and Berger (2010) and Bogdan et al. (2008), indicate that they are thinking of a marginal maximum likelihood estimate of τ. However, an empirical Bayes estimate of τ does not need to be based on this principle. Furthermore, an estimator that satisfies the second condition from Theorem 1.6 or that is truncated from below by $\frac{1}{n}$, would not be susceptible to this potential problem.

1.5 Simulation study

A simulation study provides more insight into the behaviour of the horseshoe estimator, both when using an empirical Bayes procedure with estimator (1.10) and when using the fully Bayesian procedure proposed by Carvalho et al. (2010) with a half-Cauchy prior on τ. For each data point, 100 replicates of an n-dimensional vector sampled from a $\mathcal{N}(\theta_0, I_n)$ distribution were created, where θ_0 had either 20, 40 or 200 (5%, 10% or 50%) entries equal to an integer A ranging from 1 to 10, and all the other entries equal to zero. The full Bayesian version was implemented using the code provided in (Scott, 2010), and the coordinatewise posterior mean was used as the estimator of θ_0. For the empirical Bayes procedure, the estimator (1.10) was used with $c_1 = 2$ and $c_2 = 1$. Performance was measured by squared error loss, which was averaged across replicates to create Figure 1.2.

In all settings, both estimators experience a peak in the ℓ_2 loss for values of A close to the 'universal threshold' of $\sqrt{2 \log 400} \approx 3.5$. This is not unexpected, as in the terminology of Johnstone and Silverman (2004), the horseshoe estimator is a shrinkage rule, and while it is not a thresholding rule in their sense, it does have the bounded shrinkage property which leads to thresholding-like behaviour. The bounded shrinkage property can be derived from Lemma 1.9, which yields the following inequality as τ approaches zero:

$$|T_\tau(y) - y| \leq \sqrt{2\sigma^2 \log \frac{1}{\tau}}.$$

With $\tau = \frac{1}{n}$, this leads to the 'universal threshold' of $\sqrt{2\sigma^2 \log n}$, or with $\tau = (\frac{p_n}{n})^\alpha$, a 'threshold' at $\sqrt{2\alpha\sigma^2 \log(n/p_n)}$. Based on this property and the proofs of the main results, we can divide the underlying parameters into three cases:

(i) Those that are exactly or close to zero, where the observations are shrunk close to zero;

(ii) Those that are larger than the threshold, where the horseshoe estimator essentially behaves like the identity;

(iii) Those that are close to the 'threshold', where the horseshoe estimator is most likely to shrink the observations too much.

The horseshoe estimator performs well in cases (i) and (ii) due to its pole at zero and its heavy tails respectively. The hardest parameters to recover from the noise are those that are close to the threshold, and these are the ones that affect the estimation risk the most. This phenomenon explains the peaks in the graphs of Figure 1.2 around $A = 3.5$.

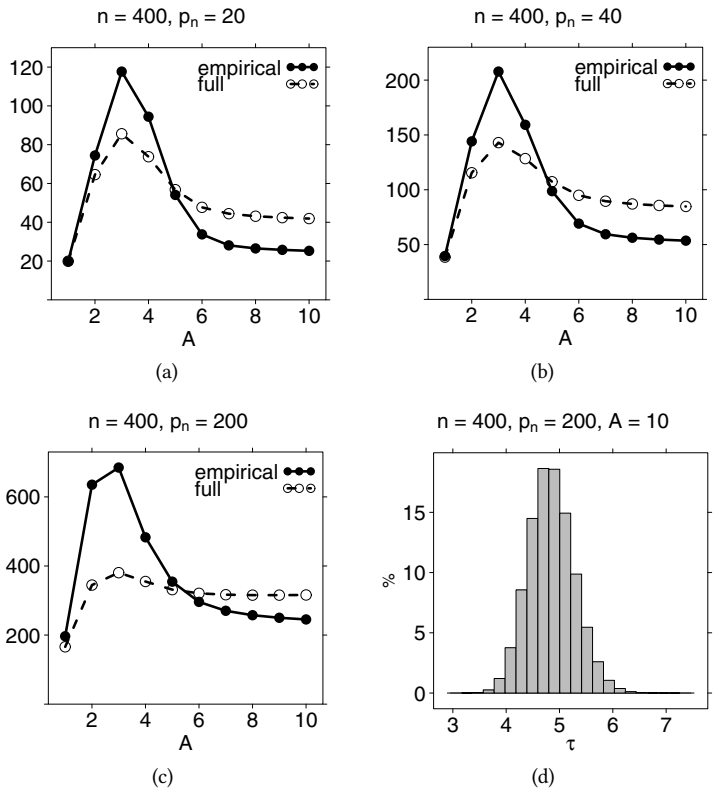

Figure 1.2: Average squared error loss over 100 replicates with underlying mean vectors of length $n = 400$ if the nonzero coefficients are taken equal to A, in case 5% (Figure (a)), 10% (Figure (b)) or 50% (Figure (c)) of the means are equal to a nonzero value A. The solid line corresponds to empirical Bayes with (1.10), $c_1 = 2, c_2 = 1$, the dashed line to full Bayes with a half-Cauchy prior on τ. Figure (d) displays a histogram of all Gibbs samples of τ (after the burn-in) of all replicates in the setting $\tau \sim C^+(0,1)$, $A = 10$, $p_n = 200$.

The full Bayes implementation with a Cauchy prior on τ attains a lower ℓ_2 loss around the universal threshold than the empirical Bayes procedure. This is because estimator (1.10) counts the number of observations that are above the universal threshold. When all the nonzero means are close to this threshold, $\widehat{\tau}$ may 'miss' some of them, thereby underestimating the sparsity level $\frac{p_n}{n}$ and thus leading to overshrinkage.

For values of A well past the universal threshold, the empirical Bayes estimator does better than the full Bayes version. For such large values of A, the estimator (1.10) will be equal to the true sparsity level with large probability and hence its good performance is not unexpected. However, an interesting question is why the full Bayes estimator does not do as well as the empirical Bayes estimator, especially because the nonzero means are so far removed from zero that the problem is 'easy'. This could be due to the choice

of a half-Cauchy prior for τ: it places no restriction on the possible values of τ and has such heavy tails that values far exceeding the sparsity level $\frac{p_n}{n}$ are possible. This would lead to undershrinkage of the observations corresponding to a zero mean, which would be reflected in the ℓ_2 loss. Figure 1.2(d) shows a histogram of all Gibbs samples of τ in the setting where 50% of the means are set equal to 10. The range of these values is (3.1, 7.3), which is very far away from $\frac{p_n}{n} = \frac{1}{2}$. This indicates that a full Bayesian version of the horseshoe prior could benefit from a different choice of prior on τ than a half-Cauchy one, for example one that is restricted to [0,1].

1.6 Concluding remarks

The choice of the global shrinkage parameter τ is critical towards ensuring the right amount of shrinkage of the observations to recover the underlying mean vector. The value of $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ was found to be optimal. Theorem 1.6 indicates that quite a wide range of estimators for τ will work well, especially in cases where the underlying mean vector is sparse. Of course, it should not come as a surprise that an estimator designed to recover sparse vectors will work especially well if the truth is indeed sparse. An interesting extension to this work would be to investigate whether the posterior concentration properties of the horseshoe prior still remain when a hyperprior is placed on τ. The result that $\tau = \frac{p_n}{n}$ (up to a log factor) yields optimal rates, together with the simulation results, suggests that in a fully Bayesian approach, a prior on τ which is restricted to $[0, 1]$ may perform better than the suggested half-Cauchy prior.

The simulation results also indicate that mean vectors with the nonzero means close to the universal threshold are the hardest to recover. In future simulations involving shrinkage rules, it would therefore be interesting to study the challenging case where all the nonzero parameters are at this threshold. The performance of the empirical Bayes estimator (1.10) leaves something to be desired around the threshold. In additional numerical experiments (not shown), we tried two other estimators of τ. The first was the 'oracle estimator' $\widehat{\tau} = \frac{p_n}{n}$. For values of the nonzero means well past the 'threshold', the behaviour of this estimator was very similar to that of (1.10). However, before the threshold, the squared error loss of the empirical procedure with the oracle estimator was between that of the full Bayes estimator and empirical Bayes with estimator (1.10). The second estimator was the mean of the samples of τ from the full Bayes estimator. The resulting squared error loss was remarkably close to that of the full Bayes estimator, for all values of the nonzero means. Neither of these two estimators is of much practical use. However, their range of behaviours suggests room for improvement over the estimator (1.10), and it would be worthwhile to study more refined estimators for τ.

An interesting question is what aspects of the horseshoe prior are truly essential towards optimal posterior contraction properties. Our proofs do not elucidate whether the pole at zero of the horseshoe prior is required, or if a prior with heavy tails, and in a sense 'sufficient' mass at zero would work as well. The failure of the Lasso to concentrate around the true mean vector at the minimax rate does indicate that heavy tails in itself may not be sufficient, and adding mass at zero solves this problem (Castillo et al., 2015; Castillo and Van der Vaart, 2012). It is possible that the pole at zero is inessential, in particular if the global tuning parameter is chosen carefully, for instance by empirical Bayes. If the tuning

parameter is chosen by a full Bayes method, the peak may be more essential, depending on its prior.

The horseshoe estimator has the property that its computational complexity depends on the sparsity level rather than the number of observations. Although there is no point mass at zero to induce sparsity, it still yields good reconstruction in ℓ_2, and a posterior distribution that contracts at an informative rate. None of the estimates will however be exactly zero. Variable selection can be performed by applying some sort of thresholding rule, such as the one suggested in (Carvalho et al., 2010) and analyzed by Datta and Ghosh (2013). The performance of this thresholding rule in simulations in the two works cited has been encouraging.

Acknowledgements

The authors would like to thank two anonymous referees for their helpful suggestions, as well as James Scott for his advice on implementing the full Bayesian version of the horseshoe estimator.

1.7 Proofs

This section begins with Lemma 1.7, providing bounds on some of the degenerate hypergeometric functions appearing in the posterior mean and posterior variance. This is followed by two lemmas that are needed for the proofs of Theorems 1.1 and 1.2: Lemma 1.8 provides two upper bounds on the horseshoe estimator and Lemma 1.9 gives a bound on the absolute value of the difference between the horseshoe estimator and an observation. We then proceed to the proof of Theorem 1.1, after which Lemma 1.10 provides upper bounds on the posterior variance. These upper bounds are then used in the proof of Theorem 1.2. The proof of Theorem 1.4 is given next, followed by Lemmas 1.11 and 1.12 supporting the proof of Theorem 1.5. This section concludes with the proofs of Theorem 1.6 and Lemma 1.13, which both concern the empirical Bayes procedure discussed in Section 1.4.

Lemma 1.7. *Define*

$$I_k(y) := \int_0^1 z^k \frac{1}{\tau^2 + (1-\tau^2)z} e^{\frac{y^2}{2\sigma^2} z} dz.$$

Then, for $a > 1$:

$$I_{\frac{3}{2}}(y) \geq \frac{1}{5}\tau^3 + \sigma^2 \frac{\tau}{y^2}\left(e^{\frac{y^2}{2a\sigma^2}} - e^{\tau^2 \frac{y^2}{2\sigma^2}}\right) + \frac{\sigma^2}{\sqrt{a}y^2}\left(e^{\frac{y^2}{2\sigma^2}} - e^{\frac{y^2}{2a\sigma^2}}\right), \tag{1.11}$$

$$I_{\frac{1}{2}}(y) \geq \frac{1}{3}\tau + \frac{\sigma^2}{y^2}\left(e^{\frac{y^2}{2\sigma^2}} - e^{\tau^2 \frac{y^2}{2\sigma^2}}\right), \tag{1.12}$$

$$I_{\frac{1}{2}}(y) \leq \frac{2}{3}e^{\tau^2 \frac{y^2}{2\sigma^2}}\tau + 2e^{\frac{y^2}{2a\sigma^2}}\left(\frac{1}{\sqrt{a}} - \tau\right) + \frac{2\sqrt{a}\sigma^2}{y^2}\left(e^{\frac{y^2}{2\sigma^2}} - e^{\frac{y^2}{2a\sigma^2}}\right), \tag{1.13}$$

$$I_{-\frac{1}{2}}(y) \geq \frac{1}{\tau} + e^{\tau^2 \frac{y^2}{2\sigma^2}}\left(\frac{1}{\tau} - \frac{1}{\sqrt{\tau}}\right) + \frac{a\sqrt{a}\sigma^2}{y^2}\left(e^{\frac{y^2}{2a\sigma^2}} - e^{\tau \frac{y^2}{2\sigma^2}}\right)$$

$$+ \frac{\sigma^2}{y^2}\left(e^{\frac{y^2}{2\sigma^2}} - e^{\frac{y^2}{2a\sigma^2}}\right), \tag{1.14}$$

$$I_{-\frac{1}{2}}(y) \leq \frac{2e^{\tau^2\frac{y^2}{2\sigma^2}}}{\tau} + 2e^{\tau\frac{y^2}{2\sigma^2}}\left(\frac{1}{\tau} - \frac{1}{\sqrt{\tau}}\right) + 2e^{\frac{y^2}{2a\sigma^2}}\left(\frac{1}{\sqrt{\tau}} - \sqrt{a}\right)$$
$$+ \frac{2a\sqrt{a}\sigma^2}{y^2}\left(e^{\frac{y^2}{2\sigma^2}} - e^{\frac{y^2}{2a\sigma^2}}\right), \tag{1.15}$$

where (1.11) *and* (1.13) *hold for* $\tau < 1/\sqrt{a}$, (1.12) *holds for* $\tau < 1$, *and* (1.14) *and* (1.15) *hold for* $\tau < 1/a$.

Proof. Write $\xi = y^2/(2\sigma^2)$. We first note that for $z \geq \tau^2$, we have $z \leq \tau^2 + (1-\tau^2)z \leq 2z$, while for $z \leq \tau^2$, we have $\tau^2 \leq \tau^2 + (1-\tau^2)z \leq 2\tau^2$. Hence, we can bound I_k from above by:

$$I_k(y) \leq \frac{1}{\tau^2}\int_0^{\tau^2} z^k e^{\xi z} dz + \int_{\tau^2}^1 z^{k-1} e^{\xi z} dz,$$

and from below by half of that quantity. We bound the integral over $[0, \tau^2]$ in all cases by bounding the factor $e^{\xi z}$ by 1 or $e^{\tau^2\xi}$. For the integral over $[\tau^2, 1]$, we first substitute $u = \xi z$, yielding: $\int_{\tau^2}^1 z^{k-1}e^{\xi z}dz = \xi^{-k}\int_{\tau^2\xi}^{\xi} u^{k-1}e^u du$. For (1.11) and (1.13), we split the domain of integration into $[\tau^2\xi, \frac{\xi}{a}]$ and $[\frac{\xi}{a}, \xi]$. For $I_{\frac{3}{2}}$, we bound by:

$$I_{\frac{3}{2}}(y) \geq \frac{1}{2}\left(\frac{1}{\tau^2}\int_0^{\tau^2} z^{\frac{3}{2}} dz + \xi^{-\frac{3}{2}}(\tau^2\xi)^{\frac{1}{2}}\int_{\tau^2\xi}^{\frac{\xi}{a}} e^u du + \xi^{-\frac{3}{2}}\left(\frac{\xi}{a}\right)^{\frac{1}{2}}\int_{\frac{\xi}{a}}^{\xi} e^u du\right),$$

yielding (1.11). Similarly, for $I_{\frac{1}{2}}$:

$$I_{\frac{1}{2}}(y) \leq \frac{1}{\tau^2}e^{\tau^2\xi}\int_0^{\tau^2} z^{\frac{1}{2}} dz + \xi^{-\frac{1}{2}}e^{\frac{\xi}{a}}\int_{\tau^2\xi}^{\frac{\xi}{a}} u^{-\frac{1}{2}} du + \xi^{-\frac{1}{2}}\left(\frac{\xi}{a}\right)^{-\frac{1}{2}}\int_{\frac{\xi}{a}}^{\xi} e^u du,$$

resulting in (1.13). The bound (1.12) is obtained similarly, but without splitting up $[\tau^2\xi, \xi]$ further, by the inequality

$$I_{\frac{1}{2}}(y) \geq \frac{1}{2\tau^2}\int_0^{\tau^2} z^{\frac{1}{2}} dz + \frac{1}{2}\xi^{-1}\int_{\tau^2\xi}^{\xi} e^u du.$$

For the bounds on $I_{-\frac{1}{2}}$, we split up the domain of integration $[\tau^2\xi, \xi]$ into $[\tau^2\xi, \tau\xi], [\tau\xi, \frac{\xi}{a}]$ and $[\frac{\xi}{a}, \xi]$, and then bound by:

$$I_{-\frac{1}{2}}(y) \geq \frac{1}{2}\left(\frac{1}{\tau^2}\int_0^{\tau^2} z^{-\frac{1}{2}} dz + \xi^{\frac{1}{2}}e^{\tau^2\xi}\int_{\tau^2\xi}^{\tau\xi} u^{-\frac{3}{2}} du + \xi^{\frac{1}{2}}\left(\frac{\xi}{a}\right)^{-\frac{3}{2}}\int_{\tau\xi}^{\frac{\xi}{a}} e^u du\right.$$
$$\left. + \xi^{\frac{1}{2}}\xi^{-\frac{3}{2}}\int_{\frac{\xi}{a}}^{\xi} e^u du\right),$$

yielding (1.14), and by:

$$I_{-\frac{1}{2}}(y) \leq \frac{1}{\tau^2} e^{\tau^2 \xi} \int_0^{\tau^2} z^{-\frac{1}{2}} dz + \xi^{\frac{1}{2}} e^{\tau \xi} \int_{\tau^2 \xi}^{\tau \xi} u^{-\frac{3}{2}} du + \xi^{\frac{1}{2}} e^{\frac{\xi}{a}} \int_{\tau \xi}^{\frac{\xi}{a}} u^{-\frac{3}{2}} du + \xi^{\frac{1}{2}} \left(\frac{\xi}{a}\right)^{-\frac{3}{2}} \int_{\frac{\xi}{a}}^{\xi} e^u du,$$

to find (1.15). □

Lemma 1.8. *If $\tau^2 < 1$, the posterior mean of the horseshoe prior can be bounded above by:*

1. *$T_\tau(y) \leq y e^{\frac{y^2}{2\sigma^2}} f(\tau)$, where f is such that $f(\tau) \leq \frac{2}{3}\tau$;*
2. $$T_\tau(y) \leq y \frac{\frac{2}{3} e^{\tau^2 \frac{y^2}{2\sigma^2}} \tau + 2 e^{\frac{y^2}{2a\sigma^2}} (\frac{1}{\sqrt{a}} - \tau) + \frac{2\sqrt{a}\sigma^2}{y^2} (e^{\frac{y^2}{2\sigma^2}} - e^{\frac{y^2}{2a\sigma^2}})}{\frac{1}{\tau} + e^{\tau^2 \frac{y^2}{2\sigma^2}} (\frac{1}{\tau} - \frac{1}{\sqrt{\tau}}) + \frac{a\sigma^2\sqrt{a}}{y^2} (e^{\frac{y^2}{2a\sigma^2}} - e^{\tau \frac{y^2}{2\sigma^2}}) + \frac{\sigma^2}{y^2} (e^{\frac{y^2}{2\sigma^2}} - e^{\frac{y^2}{2a\sigma^2}})},$$

   for any $a > 1$ and $\tau < \frac{1}{a}$.

Proof. We bound the integrals in the numerator and denominator of expression (1.1). For the first upper bound, we will use the fact that for $0 \leq z \leq 1$, $e^{\frac{y^2}{2\sigma^2} z}$ is bounded below by 1 and above by $e^{\frac{y^2}{2\sigma^2}}$. The posterior mean can therefore be bounded by:

$$T_\tau(y) \leq y e^{\frac{y^2}{2\sigma^2}} \frac{\int_0^1 z^{\frac{1}{2}} \frac{1}{\tau^2 + (1-\tau^2)z} dz}{\int_0^1 z^{-\frac{1}{2}} \frac{1}{\tau^2 + (1-\tau^2)z} dz} = y e^{\frac{y^2}{2\sigma^2}} f(\tau),$$

where

$$f(\tau) = \frac{\tau}{1-\tau^2} \left(\frac{\sqrt{1-\tau^2}}{\arctan\left(\frac{\sqrt{1-\tau^2}}{\tau}\right)} - \tau \right).$$

By Shafer's inequality for the arctangent (Shafer, 1966):

$$\frac{f(\tau)}{\tau} = \frac{1}{1-\tau^2} \left(\frac{\sqrt{1-\tau^2}}{\arctan\left(\frac{\sqrt{1-\tau^2}}{\tau}\right)} - \tau \right) < \frac{2}{3} \frac{1}{1+\tau} \leq \frac{2}{3},$$

which completes the proof for the first upper bound.

For the second inequality, we note that, in the notation of Lemma 1.7, $T_\tau(y) = y \frac{I_{\frac{1}{2}}(y)}{I_{-\frac{1}{2}}(y)}$. The bounds in Lemma 1.7 yield the stated inequality. □

Lemma 1.9. *For $\tau^2 < 1$, the absolute value of the difference between the horseshoe estimator and an observation y can be bounded by a function $h(y,\tau)$ such that for any $c > 2$:*

$$\lim_{\tau \downarrow 0} \sup_{|y| > \sqrt{c\sigma^2 \log \frac{1}{\tau}}} h(y,\tau) = 0.$$

Proof. We assume $y > 0$ without loss of generality. By a change of variables of $x = 1 - z$:

$$|T_\tau(y) - y| = y \frac{\int_0^1 e^{-\frac{y^2}{2\sigma^2}x} x(1-x)^{-\frac{1}{2}} \frac{1}{1-(1-\tau^2)x} dx}{\int_0^1 e^{-\frac{y^2}{2\sigma^2}x} (1-x)^{-\frac{1}{2}} \frac{1}{1-(1-\tau^2)x} dx}.$$

By following the proof of Watson's lemma provided in Miller (2006), we can find bounds on the numerator and denominator of the above expression. First define $g(x) = (1-x)^{-\frac{1}{2}} \frac{1}{\mathbf{1}-(1-\tau^2)x}$ and note that by Taylor's theorem, $g(x) = g(0) + xg'(\xi_x)$, where ξ_x is between 0 and x. Let s be any number between 0 and 1. Because $g''(x)$ is not negative for $x \in [0,1)$, we have that for $x \in [0,s]$, $s \in (0,1)$: $g'(0) \leq g'(x) \leq g'(s)$. The numerator can then be bounded by:

$$\begin{aligned}
\int_0^1 e^{-\frac{y^2}{2\sigma^2}x} x g(x) dx &= \int_0^s e^{-\frac{y^2}{2\sigma^2}x} x g(0) dx + \int_0^s e^{-\frac{y^2}{2\sigma^2}x} x^2 g'(\xi_x) dx \\
&\quad + \int_s^1 e^{-\frac{y^2}{2\sigma^2}x} x g(x) dx \\
&\leq \frac{1}{y^4} h_1(y,\sigma,s) + \frac{g'(s)}{y^6} h_2(y,\sigma,s) + 2e^{-\frac{sy^2}{2\sigma^2}} h_3(\tau),
\end{aligned}$$

where $h_1(y,\sigma,s) = 4\sigma^4 - 2\sigma^2(sy^2 + 2\sigma^2)e^{-\frac{sy^2}{2\sigma^2}}$, $h_2(y,\sigma,s) = 16\sigma^6 - 2\sigma^2(s^2y^4 + 4s\sigma^2y^2 + 8\sigma^4)e^{-\frac{sy^2}{2\sigma^2}}$ and $h_3(\tau) = \arctan(\frac{\sqrt{1-\tau^2}}{\tau})\tau^{-1}(1-\tau^2)^{-\frac{3}{2}} - (1-\tau^2)^{-1}$. The denominator can similarly be bounded by:

$$\begin{aligned}
\int_0^1 e^{-\frac{y^2}{2\sigma^2}x} g(x) dx &= \int_0^s e^{-\frac{y^2}{2\sigma^2}x} g(0) dx + \int_0^s e^{-\frac{y^2}{2\sigma^2}x} x g'(\xi_x) dx \\
&\quad + \int_s^1 e^{-\frac{y^2}{2\sigma^2}x} g(x) dx \\
&\geq \frac{1}{y^2} h_4(y,\sigma,s) + \frac{g'(0)}{y^4} h_5(y,\sigma,s) + 0,
\end{aligned}$$

where $h_4(y,\sigma,s) = 2\sigma^2 - 2\sigma^2 e^{-\frac{sy^2}{2\sigma^2}}$ and $h_5(y,\sigma,s) = 4\sigma^4 - 2\sigma^2 e^{-\frac{sy^2}{2\sigma^2}}(sy^2 + 2\sigma^2)$. Hence:

$$|T_\tau(y) - y| \leq \frac{\frac{1}{y} h_1(y,\sigma,s) + \frac{g'(s)}{y^3} h_2(y,\sigma,s) + 2y^3 e^{-\frac{sy^2}{2\sigma^2}} h_3(\tau)}{h_4(y,\sigma,s) + \frac{g'(0)}{y^2} h_5(y,\sigma,s)}.$$

For any fixed τ, this bound tends to zero as y tends to infinity. If $\tau \to 0$, the term containing $h_3(\tau)$ could potentially diverge. For $s = \frac{2}{3}$ and $y = \sqrt{c\sigma^2 \log(1/\tau)}$, where c is a positive

constant, this term displays the following limiting behaviour as $\tau \to 0$:

$$\lim_{\tau \downarrow 0} y^3 e^{-\frac{1}{3\sigma^2} y^2} h_3(\tau) = \lim_{\tau \downarrow 0} \left(c\sigma^2 \log \frac{1}{\tau} \right)^{\frac{3}{2}} \tau^{\frac{c}{3}-1} \left(\frac{\arctan\left(\frac{\sqrt{1-\tau^2}}{\tau} \right)}{(1-\tau^2)^{\frac{3}{2}}} - \frac{\tau}{1-\tau^2} \right)$$
$$= \begin{cases} 0 & c > 3 \\ \infty & \text{otherwise,} \end{cases}$$

because $\lim_{\tau \downarrow 0} \arctan(\frac{\sqrt{1-\tau^2}}{\tau})(1-\tau^2)^{-\frac{3}{2}} = \frac{\pi}{2}$, $\lim_{\tau \downarrow 0} \frac{\tau}{1-\tau^2} = 0$ and the factor $(c\sigma^2 \log(1/\tau))^{\frac{3}{2}} \tau^{\frac{c}{3}-1}$ tends to zero as $\tau \downarrow 0$ if $\frac{c}{3} - 1 > 0$ and infinity otherwise. The condition $c > 3$ is related to the choice of $s = \frac{2}{3}$ and can be improved to any constant strictly greater than 2 by choosing s appropriately close to one. Hence, we find that the absolute value of the difference between the posterior mean and an observation can be bounded by a function $h(y, \tau)$ with the desired property. □

Proof of Theorem 1.1

Proof. Suppose that $Y \sim \mathcal{N}(\theta, \sigma^2 I_n)$, $\theta \in \ell_0[p_n]$ and $\tilde{p}_n = \#\{i : \theta_i \neq 0\}$. Note that $\tilde{p}_n \leq p_n$. Assume without loss of generality that for $i = 1, \ldots, \tilde{p}_n$, $\theta_i \neq 0$, while for $i = \tilde{p}_n + 1, \ldots, n$, $\theta_i = 0$. We split up the expectation $\mathbb{E}_\theta \|T_\tau(Y) - \theta\|^2$ into the two corresponding parts:

$$\sum_{i=1}^{n} \mathbb{E}_{\theta_i} (T_\tau(Y_i) - \theta_i)^2 = \sum_{i=1}^{\tilde{p}_n} \mathbb{E}_{\theta_i} (T_\tau(Y_i) - \theta_i)^2 + \sum_{i=\tilde{p}_n+1}^{n} \mathbb{E}_0 T_\tau(Y_i)^2.$$

We will now show that these two terms can be bounded by $\tilde{p}_n(1 + \log \frac{1}{\tau})$ and $(n - \tilde{p}_n)\sqrt{\log(1/\tau)}\tau$ respectively, up to multiplicative constants only depending on σ, for any choice of τ such that $\tau \in (0, 1)$.

Nonzero parameters

Denote $\zeta_\tau = \sqrt{2\sigma^2 \log(1/\tau)}$. We will show

$$\mathbb{E}_{\theta_i} (T_\tau(Y_i) - \theta_i)^2 \lesssim \sigma^2 + \zeta_\tau^2. \tag{1.16}$$

for all nonzero θ_i, which can be done by bounding $\sup_y |T_\tau(y) - y|$:

$$\begin{aligned} \mathbb{E}_{\theta_i} (T_\tau(Y_i) - \theta_i)^2 &= \mathbb{E}_{\theta_i} ((T_\tau(Y_i) - Y_i) + (Y_i - \theta_i))^2 \\ &\leq 2\mathbb{E}_{\theta_i} (Y_i - \theta_i)^2 + 2\mathbb{E}_{\theta_i} (T_\tau(Y_i) - Y_i)^2 \\ &\leq 2\sigma^2 + 2 \left(\sup_y |T_\tau(y) - y| \right)^2, \end{aligned}$$

Lemma 1.9 yields the following bound on the difference between the observation and the horseshoe estimator: $|T_\tau(y) - y| \leq h(y, \tau)$, where $h(y, \tau)$ is such that $\lim_{\tau \downarrow 0} \sup_{|y| > c\zeta_\tau} h(y, \tau) = 0$ for any $c > 1$. Combining this with the inequality $|T_\tau(y) - y| \leq |y|$, we have as $\tau \to 0$:

$$\arg\max_y |T_\tau(y) - y| \lesssim \zeta_\tau, \tag{1.17}$$

which implies (1.16), as $|T_\tau(y)| \leq |y|$:

$$\left(\sup_y |T_\tau(y) - y|\right)^2 \lesssim \zeta_\tau^2.$$

Parameters equal to zero

We split up the term for the zero means into two parts:

$$\mathbb{E}_0 T_\tau(Y)^2 = \mathbb{E}_0 T_\tau(Y)^2 \mathbf{1}_{|Y| \leq \zeta_\tau} + \mathbb{E}_0 T_\tau(Y)^2 \mathbf{1}_{|Y| > \zeta_\tau},$$

where $\zeta_\tau = \sqrt{2\sigma^2 \log(1/\tau)}$. For the first term, we have, by the first bound in Lemma 1.8:

$$\begin{aligned}
\mathbb{E}_0 T_\tau(Y)^2 \mathbf{1}_{\{|Y| \leq \zeta_\tau\}} &= \int_{-\zeta_\tau}^{\zeta_\tau} T_\tau(y)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \\
&\leq \int_{-\zeta_\tau}^{\zeta_\tau} y^2 e^{\frac{y^2}{\sigma^2}} f(\tau)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy = \frac{f(\tau)^2}{\sqrt{2\pi\sigma^2}} \int_{-\zeta_\tau}^{\zeta_\tau} y^2 e^{\frac{y^2}{2\sigma^2}} dy \\
&\leq \sqrt{\frac{2}{\pi}} \sigma f(\tau)^2 \zeta_\tau \frac{1}{\tau} \leq \sqrt{\frac{2}{\pi}} \sigma \frac{4}{9} \zeta_\tau \tau \lesssim \zeta_\tau \tau,
\end{aligned}$$

where the identity $\frac{d}{dy} y e^{\frac{y^2}{2\sigma^2}} = \frac{y^2}{\sigma^2} e^{\frac{y^2}{2\sigma^2}} + e^{\frac{y^2}{2\sigma^2}}$ was used to bound $\int_{-\zeta_\tau}^{\zeta_\tau} y^2 e^{\frac{y^2}{2\sigma^2}} dy$. For the second term, because $|T_\tau(y)| \leq |y|$ for all y, we have by the identity $y^2\phi(y) = \phi(y) - \frac{d}{dy}[y\phi(y)]$, and by Mills' ratio:

$$\begin{aligned}
\mathbb{E}_0 T_\tau(Y)^2 \mathbf{1}_{\{|Y| > \zeta_\tau\}} &\leq \mathbb{E}_0 Y^2 \mathbf{1}_{\{|Y| > \zeta_\tau\}} = 2 \int_{\frac{\zeta_\tau}{\sigma}}^{\infty} \sigma^2 y^2 \phi(y) dy \\
&\leq 2\sigma\zeta_\tau \phi\left(\frac{\zeta_\tau}{\sigma}\right) + 2\sigma^3 \frac{\phi\left(\frac{\zeta_\tau}{\sigma}\right)}{\zeta_\tau} \leq 4\sigma\zeta_\tau \phi\left(\frac{\zeta_\tau}{\sigma}\right) = 4\sigma\zeta_\tau \frac{1}{\sqrt{2\pi}} \tau,
\end{aligned}$$

where the last inequality holds for $\zeta_\tau > \sigma^2$. If we apply this inequality and combine this upper bound with the upper bound on the first term, we find, for $\zeta_\tau > \sigma^2$ (corresponding to $\tau < e^{-\frac{\sigma^2}{2}}$):

$$\mathbb{E}_0 T_\tau(Y)^2 = \mathbb{E}_0 T_\tau(Y)^2 \mathbf{1}_{\{|Y| \leq \zeta_\tau\}} + \mathbb{E}_0 T_\tau(Y)^2 \mathbf{1}_{\{|Y| > \zeta_\tau\}} \lesssim \zeta_\tau \tau. \tag{1.18}$$

Hence, for $\tau < e^{-\frac{\sigma^2}{2}}$:

$$\sum_{i=p_n+1}^{n} \mathbb{E}_0 T_\tau(Y_i)^2 \lesssim (n - p_n)\zeta_\tau \tau. \tag{1.19}$$

Conclusion

By (1.16) and (1.19), we find for $\tau < e^{-\frac{\sigma^2}{2}}$:

$$\sum_{i=1}^{n} \mathbb{E}_{\theta_i}(T_\tau(Y_i) - \theta_i)^2 \lesssim \tilde{p}_n(1 + \zeta_\tau^2) + (n - \tilde{p}_n)\tau\zeta_\tau. \qquad \square$$

Lemma 1.10. *The posterior variance when using the horseshoe prior can be expressed as:*

$$\operatorname{var}(\theta \mid y) = \frac{\sigma^2}{y} T_\tau(y) - (T_\tau(y) - y)^2 + y^2 \frac{\displaystyle\int_0^1 (1-z)^2 z^{-\frac{1}{2}} \frac{1}{\tau^2 + (1-\tau^2)z} e^{\frac{y^2}{2\sigma^2} z} dz}{\displaystyle\int_0^1 z^{-\frac{1}{2}} \frac{1}{\tau^2 + (1-\tau^2)z} e^{\frac{y^2}{2\sigma^2} z} dz}, \quad (1.20)$$

and bounded from above by:

1. $\operatorname{var}(\theta \mid y) \leq \sigma^2 + y^2$;
2. $\operatorname{var}(\theta \mid y) \leq (\frac{\sigma^2}{y} + y) T_\tau(y) - T_\tau(y)^2$.

Proof. As proven in Pericchi and Smith (1992):

$$\operatorname{var}(\theta \mid y) = \sigma^2 + \sigma^4 \frac{d^2}{dy^2} \log m(y) = \sigma^2 - \left(\sigma^2 \frac{m'(y)}{m(y)} \right)^2 + \sigma^4 \frac{m''(y)}{m(y)},$$

where $m(y)$ is the density of the marginal distribution of y. Equality (1.20) can be found by combining the expressions

$$m(y) = \frac{1}{\sqrt{2\pi^3}\sigma\tau} e^{-\frac{y^2}{2\sigma^2}} \int_0^1 z^{-\frac{1}{2}} \frac{1}{1 - \left(1 - \frac{1^2}{\tau^2}\right) z} e^{\frac{y^2}{2\sigma^2} z} dz$$

$$m''(y) = \frac{1}{y} m'(y) + \frac{1}{\sqrt{2\pi^3}\sigma\tau} \frac{y^2}{\sigma^4} e^{-\frac{y^2}{2\sigma^2}} \int_0^1 z^{-\frac{1}{2}} (1-z)^2 \frac{1}{1 - \left(1 - \frac{1}{\tau^2}\right) z} e^{\frac{y^2}{2\sigma^2} z} dz$$

with the equality $T_\tau(y) = y + \sigma^2 \frac{m'(y)}{m(y)}$. The first upper bound is implied by the property $|T_\tau(y)| < |y|$ and the fact that $(1-z)^2 \leq 1$ for $z \in [0,1]$. The second upper bound can be demonstrated by noting that $(1-z)^2 \leq 1 - z$ for $z \in [0,1]$ and hence:

$$\operatorname{var}(\theta \mid y) \leq \frac{\sigma^2}{y} T_\tau(y) - (y - T_\tau(y))^2 + y^2 \left(1 - \frac{1}{y} T_\tau(y)\right). \qquad \square$$

Proof of Theorem 1.2

Proof. As in the proof of Theorem 1.1 we assume that $\theta_i \neq 0$ for $i = 1, \ldots, \tilde{p}_n$ and $\theta_i = 0$ for $i = \tilde{p}_n + 1, \ldots, n$, where $\tilde{p}_n \leq p_n$ by assumption. We consider the posterior variances for the zero and nonzero means separately. Denote $\zeta_\tau = \sqrt{2\sigma^2 \log(1/\tau)}$.

Nonzero means

By applying the same reasoning as in Lemma 1.9 to the final term of $\operatorname{var}(\theta|y)$ in (1.20), we can find a function $\tilde{h}(y,t)$ such that $\operatorname{var}(\theta|y) \leq \tilde{h}(y,\tau)$, where $\tilde{h}(y,\tau) \to \sigma^2$ as $y \to \infty$ for any fixed τ. If $\tau \to 0$, the function $\tilde{h}(y,\tau)$ displays the following limiting behaviour for any $c > 1$:

$$\lim_{\tau \downarrow 0} \sup_{|y| > c\zeta_\tau} \tilde{h}(y,\tau) = \sigma^2.$$

Hence, as $\tau \to 0$: $\text{var}(\theta|y) \lesssim \sigma^2$, for any $|y|$ that increases as least as fast as ζ_τ when τ decreases. Now suppose $|y| \leq \zeta_\tau$. Then, by the bound $\text{var}(\theta \mid y) \leq \sigma^2 + y^2$ from Lemma 1.10, we find:

$$\text{var}(\theta \mid y) \leq \sigma^2 + \zeta_\tau^2.$$

Therefore:

$$\sum_{i=1}^{\tilde{p}_n} \mathbb{E}_{\theta_i} \text{var}(\theta_i \mid Y_i) \lesssim \tilde{p}_n(1 + \zeta_\tau^2). \tag{1.21}$$

Zero means

By the bound $\text{var}(\theta \mid y) \leq \sigma^2 + y^2$, we find for $c \geq 1$:

$$\begin{aligned}
\mathbb{E}_0 \,\text{var}(\theta \mid Y)\mathbf{1}_{\{|Y|>c\zeta_\tau\}} &\leq 2\int_{c\zeta_\tau}^{\infty} (\sigma^2 + y^2)\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \\
&= 2\sigma^2\Phi^c\left(\frac{c\zeta_\tau}{\sigma}\right) + 2\int_{\frac{c\zeta_\tau}{\sigma}}^{\infty} \sigma^2 x^2 \phi(x) dx \\
&\leq 4\sigma^3 \frac{\phi\left(\frac{c\zeta_\tau}{\sigma}\right)}{c\zeta_\tau} + 2\sigma c\zeta_\tau \phi\left(\frac{c\zeta_\tau}{\sigma}\right) \lesssim \frac{\tau}{\zeta_\tau} + \zeta_\tau\tau.
\end{aligned}$$

For $|y| < c\zeta_\tau$, we consider the upper bound $\text{var}(\theta \mid y) \leq (\frac{\sigma^2}{y} + y)T_\tau(y) - T_\tau(y)^2$ from Lemma 1.10. From this bound, we get $\text{var}(\theta \mid y) \leq \frac{\sigma^2}{y}T_\tau(y) + yT_\tau(y)$. Hence:

$$\begin{aligned}
\mathbb{E}_0 \,\text{var}(\theta \mid Y)\mathbf{1}_{\{|Y|\leq c\zeta_\tau\}} \leq \sigma^2 &\int_{-c\zeta_\tau}^{c\zeta_\tau} \frac{1}{y} T_\tau(y) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \\
&+ \int_{-c\zeta_\tau}^{c\zeta_\tau} yT_\tau(y) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy.
\end{aligned} \tag{1.22}$$

We bound the first integral from (1.22) by applying the first bound on $T_\tau(y)$ from Lemma 1.8:

$$\begin{aligned}
\sigma^2 \int_{-c\zeta_\tau}^{c\zeta_\tau} \frac{1}{y} T_\tau(y) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy &\leq \sigma^2 \int_{-c\zeta_\tau}^{c\zeta_\tau} f(\tau) \frac{1}{\sqrt{2\pi\sigma^2}} dy \\
&= \sqrt{\frac{2\sigma}{\pi}} c\zeta_\tau f(\tau) \lesssim \zeta_\tau\tau,
\end{aligned}$$

because $f(\tau) \leq \frac{2}{3}\tau$. For the second term in (1.22), we first note that the second bound from Lemma 1.8 can be relaxed to:

$$T_\tau(y) \leq \tau y\left(\frac{2}{3}\tau e^{\tau^2\frac{y^2}{2\sigma^2}} + \frac{2}{\sqrt{a}} e^{\frac{y^2}{2a\sigma^2}} + 2\sqrt{a}\sigma^2 \frac{1}{y^2} e^{\frac{y^2}{2\sigma^2}}\right) \tag{1.23}$$

for any $a > 1$ and $\tau < \frac{1}{a}$. By plugging this bound into the second integral of (1.22), we get three terms, which we will name I_1, I_2 and I_3 respectively. We then find, bounding above by the integral over $\mathbb{R}$ instead of $[-c\zeta_\tau, c\zeta_\tau]$ for I_1 and I_2:

$$I_1 = \frac{2}{3}\tau^2 \int_{-c\zeta_\tau}^{c\zeta_\tau} y^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1-\tau^2)\frac{y^2}{2\sigma^2}} dy \leq \frac{2}{3}\tau^2 \frac{\sigma^2}{(1-\tau^2)^{\frac{3}{2}}} \lesssim \tau^2.$$

$$I_2 = \frac{2}{\sqrt{a}}\tau \int_{-c\zeta_\tau}^{c\zeta_\tau} y^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{a-1}{a}\frac{y^2}{2\sigma^2}} dy \leq \frac{2a\sigma^2}{(a-1)^{\frac{3}{2}}}\tau \lesssim \tau.$$

$$I_3 = 2\sqrt{a}\sigma^2\tau \int_{-c\zeta_\tau}^{c\zeta_\tau} \frac{1}{\sqrt{2\pi\sigma^2}} dy = \frac{2\sqrt{2a}c\sigma}{\sqrt{\pi}}\zeta_\tau\tau \lesssim \zeta_\tau\tau.$$

And thus:

$$\sum_{i=\tilde{p}_n+1}^{n} \mathbb{E}_0 \operatorname{var}(\theta_i \mid Y_i) \lesssim (n - \tilde{p}_n)\,(\zeta_\tau + \tau + 1)\,\tau. \tag{1.24}$$

Conclusion

By (1.21) and (1.24):

$$\mathbb{E}_\theta \sum_{i=1}^{n} \operatorname{var}(\theta_i \mid Y_i) \lesssim \tilde{p}_n(1 + \zeta_\tau^2) + (n - \tilde{p}_n)\,(\zeta_\tau + \tau + 1)\,\tau. \qquad \square$$

Proof of Theorem 1.4

Proof. By expanding $(1-z)^2 z^{-\frac{1}{2}} = z^{-\frac{1}{2}} - 2z^{\frac{1}{2}} + z^{\frac{3}{2}}$, we see that the final term in (1.20) is equal to:

$$y^2 - 2yT_\tau(y) + y^2 \frac{\int_0^1 z^{\frac{3}{2}} \frac{1}{\tau^2+(1-\tau^2)z} e^{\frac{y^2}{2\sigma^2}z} dz}{\int_0^1 z^{-\frac{1}{2}} \frac{1}{\tau^2+(1-\tau^2)z} e^{\frac{y^2}{2\sigma^2}z} dz}.$$

As $\frac{T_\tau(y)}{y}$ is non-negative, we can bound the posterior variance from below by the final two terms in (1.20). By the above equality, this yields the following lower bound:

$$\operatorname{var}(\theta \mid y) \geq y^2 \frac{I_{\frac{3}{2}}(y)}{I_{-\frac{1}{2}}(y)} - T_\tau(y)^2 = y^2 \left(\frac{I_{\frac{3}{2}}(y)}{I_{-\frac{1}{2}}(y)} - \left(\frac{I_{\frac{1}{2}}(y)}{I_{-\frac{1}{2}}(y)} \right)^2 \right),$$

where I_k is as in Lemma 1.7. We now use the bounds from Lemma 1.7 with $a = 2$ and take ξ equal to $c\log(1/\tau)$ for some nonnegative constant c. Then $e^{\xi} = \frac{1}{\tau^c}$ and $e^{\frac{\xi}{2}} = \frac{1}{\tau^{\frac{c}{2}}}$. Taking for each bound on I_k, $k \in \{\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}\}$, the term that diverges fastest as τ approaches zero, we find that the lower bound is asymptotically of the order:

$$2\sigma^2\xi \left(\frac{\frac{1}{2\sqrt{2}\xi}\frac{1}{\tau^c}}{\max\left\{\frac{2e^{\tau\xi}}{\tau}, \frac{2\sqrt{2}}{\xi}\frac{1}{\tau^c}\right\}} - \left(\frac{\frac{\sqrt{2}}{\xi}\frac{1}{\tau^c}}{\max\{\frac{e^{\tau^2\xi}}{\tau}, \frac{1}{2\xi}\frac{1}{\tau^c}\}} \right)^2 \right).$$

For $c \leq 1$, this reduces to:

$$\frac{\sigma^2}{2\sqrt{2}} e^{-\tau\xi}\tau^{1-c} - \frac{4\sigma^2}{\xi} e^{-2\tau^2\xi}\tau^{2-2c}.$$

The second term is negligible compared to the first. Hence, we will use the term $\frac{\sigma^2}{2\sqrt{2}} e^{-\tau\xi}\tau^{1-c}$ as our lower bound on $\operatorname{var}(\theta \mid y)$ for $y = \pm\sqrt{2c\sigma^2\log(1/\tau)} = \sqrt{c}\zeta_\tau$, where $\zeta_\tau = \sqrt{2\sigma^2\log(1/\tau)}$. To find the lower bound on $\sum_{i=1}^n \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid Y_i)$, we only need to consider the parameters equal to zero:

$$\sum_{i=1}^{n} \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid Y_i) \geq (n - p_n)\mathbb{E}_0 \operatorname{var}(\theta_i \mid Y_i)\mathbf{1}_{\{|Y_i| \leq \zeta_\tau\}}. \tag{1.25}$$

By the substitution $x = y^2/\zeta_\tau^2, dy = \frac{\sigma\sqrt{\log(1/\tau)}}{\sqrt{2x}}dx$, we find:

$$\begin{aligned} E_0 \operatorname{var}(\theta_i \mid Y_i)\mathbf{1}_{\{|Y_i| \leq \zeta_\tau\}} &\geq 2\int_0^{\zeta_\tau} \frac{\sigma^2}{2\sqrt{2}} e^{-\tau\frac{y^2}{2\sigma^2}} \tau^{1-\frac{y^2}{\zeta_\tau^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \\ &= \frac{\sigma}{4\sqrt{\pi}}\tau\zeta_\tau \int_0^1 \frac{\tau^{\tau x}}{\sqrt{x}} dx \geq \frac{\sigma}{2\sqrt{\pi}} e^{-\frac{1}{e}}\tau\zeta_\tau, \end{aligned} \tag{1.26}$$

where in the last step, we used $\tau^{\tau x} \geq \tau^\tau \geq e^{-\frac{1}{e}}$ for $x \in [0,1], \tau \in (0,1]$. By plugging this into (1.25), we find that as $\tau \to 0$:

$$\sum_{i=1}^{n} \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid Y_i) \gtrsim (n - p_n)\tau\zeta_\tau, \tag{1.27}$$

finishing the proof for the first statement of the theorem.

We now consider θ such that $\theta_i = a_n$ for $i = 1, \ldots, p_n$, and $\theta_i = 0$ for $i = p_n + 1, \ldots, n$, and assume without loss of generality that $a_n > 0$. We wish to find conditions on a_n such that the lower bound (1.27) is sharp (up to a constant factor). Denoting $\zeta_\tau = \sqrt{2\sigma^2 \log(1/\tau)}$, as before, it is sufficient if we can find a_n such that $\mathbb{E}_{\theta_i = a_n} \operatorname{var}(\theta_i \mid Y_i) \lesssim \tau\zeta_\tau$, because in combination with the bound (1.24), this will yield $\sum_{i=1}^n \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid Y_i) \lesssim n\tau\zeta_\tau$, which is of the same order as (1.27), as $p_n = o(n)$. Sufficient conditions on a_n can be found by adapting the proof for the 'zero means' case of Theorem 1.2.

We first consider $|y_i| > \zeta_\tau$. By the first bound of Lemma 1.10:

$$\begin{aligned} \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid Y_i)\mathbf{1}_{\{|Y_i| > \zeta_\tau\}} \leq &\int_{\zeta_\tau}^{\infty} (\sigma^2 + y^2) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy \\ &+ \int_{-\infty}^{-\zeta_\tau} (\sigma^2 + y^2) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy. \end{aligned} \tag{1.28}$$

The first integral from (1.28) can be split into two parts by splitting up the factor $\sigma^2 + y^2$, the first of which can be bounded, by substituting $x = (y - a_n)/\sigma$ and applying Mills' ratio:

$$\sigma^2 \int_{(\zeta_\tau - a_n)/\sigma}^{\infty} \phi(x)dx = \sigma^2 \Phi^c\left(\frac{\zeta_\tau - a_n}{\sigma}\right) \leq \frac{\sigma^3}{\zeta_\tau - a_n}\phi\left(\frac{\zeta_\tau - a_n}{\sigma}\right). \tag{1.29}$$

The second of these integrals is, by $y^2 = (y - a_n)^2 - a_n^2 + 2a_n y$, equal to:

$$\begin{aligned} &\int_{\zeta_\tau}^{\infty} (y - a_n)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy - a_n^2 \int_{\zeta_\tau}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy \\ &\quad + a_n \int_{\zeta_\tau}^{\infty} y \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy. \end{aligned} \tag{1.30}$$

The second integral of (1.30) can be bounded from below by zero, and the third from above by $a_n \mathbb{E}_{\theta_i} Y_i = a_n^2$. Again substituting $x = (y - a_n)/\sigma$ yields the following upper bound on (1.30): $\sigma^2 \int_{(\zeta_\tau - a_n)/\sigma}^{\infty} x^2 \phi(x) dx + a_n^2$. Now using the equality $x^2\phi(x) = \phi(x) - \frac{d}{dx}[x\phi(x)]$ and again Mills' ratio, and combining with (1.29), we find the following upper bound on the first integral from (1.28):

$$\frac{2\sigma^3}{\zeta_\tau - a_n}\phi\left(\frac{\zeta_\tau - a_n}{\sigma}\right) + \sigma(\zeta_\tau - a_n)\phi\left(\frac{\zeta_\tau - a_n}{\sigma}\right) + a_n^2. \tag{1.31}$$

By substituting $x = -y$ in the second integral from (1.28) and then applying the same inequalities to it as to the first integral, the following bound is obtained:

$$\frac{2\sigma^3}{\zeta_\tau + a_n}\phi\left(\frac{\zeta_\tau + a_n}{\sigma}\right) + \sigma(\zeta_\tau + a_n)\phi\left(\frac{\zeta_\tau + a_n}{\sigma}\right). \tag{1.32}$$

This bound does not include a term a_n^2, because in the step equivalent to (1.30), the identity $y^2 = (y + a_n)^2 - a_n^2 - 2ya_n$ is used, and thus only the integral $\int_{\zeta_\tau}^{\infty}(y + a_n)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y+a_n)^2}{2\sigma^2}} dy$ needs to be bounded in that step. $\mathbb{E}_{\theta_i} \operatorname{var}(\theta \mid Y)\mathbf{1}_{\{|Y|>\zeta_\tau\}}$ can thus be bounded by the sum of (1.31) and (1.32). The factor $\phi((\zeta_\tau + a_n)/\sigma)$ can be bounded from above by $\phi(\zeta_\tau/\sigma) = \tau/\sqrt{2\pi}$. The factor $\phi((\zeta_\tau - a_n)/\sigma)$ is equal to $\frac{1}{\sqrt{2\pi}} e^{-\frac{\zeta_\tau^2}{2\sigma^2}} e^{-\frac{a_n^2}{2\sigma^2}} e^{\frac{\zeta_\tau a_n}{\sigma}} = \frac{\tau}{\sqrt{2\pi}} e^{-\frac{a_n^2}{2\sigma^2}} e^{\frac{\zeta_\tau a_n}{\sigma}}$. Hence we arrive at the following upper bound:

$$\frac{\sigma}{\sqrt{2\pi}}\left[\left(\frac{2\sigma^2}{\zeta_\tau - a_n} + \zeta_\tau - a_n\right) e^{-\frac{a_n^2}{2\sigma^2}} e^{\frac{\zeta_\tau a_n}{\sigma}} + \frac{2\sigma^2}{\zeta_\tau + a_n} + \zeta_\tau + a_n\right]\tau + a_n^2. \tag{1.33}$$

If $a_n \lesssim 1/\zeta_\tau$, then $e^{-\frac{a_n^2}{2\sigma^2}} e^{\frac{\zeta_\tau a_n}{\sigma}} = O(1)$ and $\zeta_\tau \pm a_n = O(\zeta_\tau)$, yielding an upper bound on (1.33) of order $\tau\zeta_\tau$.

We now consider $|y_i| \le \zeta_\tau$. We use the second bound of Lemma 1.10:

$$\begin{aligned}\mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid Y_i)\mathbf{1}_{\{|Y_i|\le\zeta_\tau\}} \le{}& \sigma^2 \int_{-\zeta_\tau}^{\zeta_\tau} \frac{1}{y} T_\tau(y) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy \\ &+ \sigma^2 \int_{-\zeta_\tau}^{\zeta_\tau} y T_\tau(y) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy.\end{aligned} \tag{1.34}$$

Applying inequality $\frac{1}{y}T_\tau(y) \le \frac{2}{3}\tau e^{\frac{y^2}{2\sigma^2}}$ from Lemma 1.8 to the first integral yields the bound:

$$\frac{\sqrt{2}\sigma}{3\sqrt{\pi}}\tau \int_{-\zeta_\tau}^{\zeta_\tau} e^{\frac{y^2}{2\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy = \frac{\sqrt{2}\sigma}{3\sqrt{\pi}}\tau e^{-\frac{a_n^2}{2\sigma^2}} \int_{-\zeta_\tau}^{\zeta_\tau} e^{\frac{a_n y}{\sigma^2}} dy \le \frac{\sqrt{2}\sigma}{3\sqrt{\pi}}\tau e^{-\frac{a_n^2}{2\sigma^2}} 2\zeta_\tau e^{\frac{a_n\zeta_\tau}{\sigma^2}}.$$

If $a_n \lesssim 1/\zeta_\tau$, we have $a_n\zeta_\tau = O(1)$ and thus this term will be of order $\tau\zeta_\tau$. For the second integral from (1.34), we use bound (1.23). This leads to three integrals to be bounded, I_1, I_2 en I_3.

$$I_1 = \frac{\sigma}{\sqrt{2\pi}}\frac{2}{3}\tau^2 e^{\frac{\tau^2}{1-\tau^2}\frac{a_n^2}{2\sigma^2}} \int_{-\zeta_\tau}^{\zeta_\tau} y^2 e^{-\frac{1}{2\sigma^2/(1-\tau^2)}\left(y - \frac{a_n}{1-\tau^2}\right)^2} dy$$

$$\leq \frac{2}{3} e^{\frac{\tau^2}{1-\tau^2}\frac{a_n^2}{2\sigma^2}} \frac{\sigma^2}{(1-\tau^2)^{3/2}} \left(\sigma^2 + \frac{a_n^2}{1-\tau^2}\right) \tau^2.$$

$$I_2 = \frac{2\sigma}{\sqrt{a}\sqrt{2\pi}} \tau e^{\frac{a_n^2}{(b-1)2\sigma^2}} \int_{-\zeta_\tau}^{\zeta_\tau} y^2 e^{-\frac{1}{2\sigma^2}\frac{a}{a-1}\left(y - \frac{a}{a-1}a_n\right)^2}$$

$$\leq \frac{2}{\sqrt{b}} e^{\frac{a_n^2}{(b-1)2\sigma^2}} \sigma^2 \left(\frac{a}{a-1}\right)^{3/2} \left(\sigma^2 + \frac{a}{a-1}a_n^2\right) \tau.$$

$$I_3 = \frac{2\sqrt{a}\sigma^3}{\sqrt{2\pi}} \tau \int_{-\zeta_\tau}^{\zeta_\tau} e^{\frac{y^2}{2\sigma^2}} e^{-\frac{(y-a_n)^2}{2\sigma^2}} dy \leq \frac{2\sqrt{2a}\sigma^3}{\sqrt{\pi}} e^{-\frac{a_n^2}{2\sigma^2}} e^{\frac{a_n\zeta_\tau}{\sigma^2}} \tau\zeta_\tau.$$

I_1, I_2 and I_3 will all be of no larger order than $\tau\zeta_\tau$ if $a_n \lesssim 1/\zeta_\tau$. □

Lemma 1.11. *For all $k \in \mathbb{R}$, $\int_1^y u^k e^u du = y^k e^y (1 + O(1/y))$, as $y \to \infty$.*

Proof. For $k = 0$, the statement is immediate. By integration by parts the integral is seen to be equal to $y^k e^y - e - \int_1^y k u^{k-1} e^u du$. For $k \neq 0$, the latter integral is bounded above by

$$|k| \int_1^{y/2} (1 \vee y/2)^{k-1} e^u du + |k| \int_{y/2}^{y} (y/2 \vee y)^{k-1} e^u du.$$

This is further bounded above by a multiple of $(1 \vee y^{k-1})e^{y/2} + y^{k-1}e^y$. □

Lemma 1.12. *Let I_k be as in Lemma 1.7. There exist functions R_k with $\sup_{\zeta_\tau/4 \leq y \leq 4\zeta_\tau} |R_k(y)| \to 0$ for $k > 0$ and $k = -\frac{1}{2}$, such that,*

$$I_k(y) = \left(\tau^{2k} \int_0^1 \frac{z^k}{1+z} dz + \frac{2\sigma^2}{y^2} e^{\frac{y^2}{2\sigma^2}}\right) (1 + R_k(y)), \quad \text{for } k > 0,$$

$$I_{-\frac{1}{2}}(y) = \left(\tau^{-1} \int_0^\infty \frac{1}{\sqrt{z}(1+z)} dz + \frac{2\sigma^2}{y^2} e^{\frac{y^2}{2\sigma^2}}\right) \left(1 + R_{-\frac{1}{2}}(y)\right).$$

Proof. We split the integral in the definition of I_k over the intervals $[0, \tau^2]$ and $[\tau^2, 1]$. The first interval contributes, uniformly in $y\tau \to 0$,

$$\int_0^{\tau^2} \frac{z^k e^{\frac{y^2}{2\sigma^2}z}}{\tau^2 + (1-\tau^2)z} dz = \int_0^{\tau^2} \frac{z^k}{\tau^2 + (1-\tau^2)z} dz \, (1 + o(1))$$

$$= \tau^{2k} \int_0^1 \frac{u^k}{1 + (1-\tau^2)u} du \, (1 + o(1)), \tag{1.35}$$

by the substitution $u = z/\tau^2$. The integral tends to $\int_0^1 \frac{u^k}{1+u} du$, by the dominated convergence theorem, for any $k > -1$. The second interval contributes, with the substitution $u = (y^2/2\sigma^2)z$:

$$\int_{\tau^2}^1 \frac{z^k e^{\frac{y^2}{2\sigma^2}z}}{\tau^2 + (1-\tau^2)z} dz = \left(\frac{2\sigma^2}{y^2}\right)^k \left(\int_{\frac{y^2}{2\sigma^2}\tau^2}^1 + \int_1^{\frac{y^2}{2\sigma^2}}\right) \frac{u^k e^u}{\frac{y^2}{2\sigma^2}\tau^2 + (1-\tau^2)u} du.$$

In the second integral the argument satisfies $u \geq 1$, and hence $u/((y^2\tau^2/(2\sigma^2)+(1-\tau^2)) \to 1$, uniformly in u and $y\tau \to 0$. Hence

$$\left(\frac{2\sigma^2}{y^2}\right)^k \int_1^{\frac{y^2}{2\sigma^2}} \frac{u^k e^u}{\frac{y^2}{2\sigma^2}\tau^2 + (1-\tau^2)u} du \asymp \left(\frac{2\sigma^2}{y^2}\right)^k \int_1^{\frac{y^2}{2\sigma^2}} u^{k-1} e^u du$$
$$\asymp \frac{2\sigma^2}{y^2} e^{\frac{y^2}{2\sigma^2}} (1 + o(1))$$

as $y \to \infty$, by Lemma 1.11. For the first integral we separately consider the cases $k > 0$ and $k = -1/2$. If $k > 0$, then $\int_0^1 u^{k-1} e^u du$ converges, and hence, by the dominated convergence theorem, uniformly in $y\tau \to 0$,

$$\left(\frac{2\sigma^2}{y^2}\right)^k \int_{\tau^2 \frac{y^2}{2\sigma^2}}^1 \frac{u^k e^u}{\frac{y^2}{2\sigma^2}\tau^2 + (1-\tau^2)u} du \to \left(\frac{2\sigma^2}{y^2}\right)^k \int_0^1 u^{k-1} e^u du.$$

If $k = -1/2$, then we substitute $v = 2\sigma^2 u/(\tau^2 y^2)$ and rewrite the integral as

$$\left(\frac{2\sigma^2}{y^2}\right)^{-\frac{1}{2}} \int_1^{\frac{2\sigma^2}{\tau^2 y^2}} \frac{v^{-\frac{1}{2}} e^{\frac{\tau^2 y^2}{2\sigma^2} v}}{1 + (1-\tau^2)v} \left(\frac{\tau^2 y^2}{2\sigma^2}\right)^{-\frac{1}{2}} dv = \frac{1}{\tau} \int_1^\infty \frac{v^{-1/2}}{1+v} dv (1 + o(1)).$$

This combines with the integral (1.35). □

Proof of Theorem 1.5

Proof. Denote $\zeta_\tau = \sqrt{2\sigma^2 \log(1/\tau)}$ and assume that $\theta_i = \gamma\zeta_\tau$ for $i = 1, \ldots, p_n$ and $\theta_i = 0$ for $i = p_n + 1, \ldots, n$. We prove (1.7) by proving that there exists a positive constant $c_1(\gamma)$ such that

$$\mathbb{E}_{\theta=\gamma\zeta_\tau} T_\tau(Y) = \tau^{(1-\gamma)^2} \zeta_\tau^{2\gamma-2} c_1(\gamma)(1 + o(1)). \tag{1.36}$$

If (1.36) holds, we have, by Jensen's inequality:

$$\sum_{i=1}^{p_n} \mathbb{E}_{\theta_i} (T_\tau(Y_i) - \theta_i)^2 \geq p_n (\tau^{(1-\gamma)^2} \zeta_\tau^{2\gamma-2} c_1(\gamma) - \gamma\zeta_\tau)^2 \gtrsim p_n \zeta_\tau^2, \tag{1.37}$$

as $\tau \to 0$. In addition, we have $T_\tau(y) = y I_{\frac{1}{2}}(y)/I_{-\frac{1}{2}}(y)$. For $|y| = \sqrt{2\sigma^2 c \log(1/\tau)}$, with $c > 1$, the lower bound (1.12) on $I_{\frac{1}{2}}(y)$ behaves as $(\sigma^2/y^2) e^{\frac{y^2}{2\sigma^2}}$, while the upper bound (1.15) on $I_{-\frac{1}{2}}(y)$ behaves as $(2a\sqrt{a}\sigma^2/y^2) e^{\frac{y^2}{2\sigma^2}}$, as $\tau \to 0$. Therefore, for $|y| > \zeta_\tau$, we have $T_\tau(y) \gtrsim y$. Thus, we can bound by:

$$\sum_{i=p_n+1}^{n} \mathbb{E}_{\theta_i} T_\tau(Y_i)^2 \geq (n - p_n) \mathbb{E}_{\theta=0} T_\tau(Y)^2 \mathbf{1}_{\{|Y| > \zeta_\tau\}} \gtrsim (n - p_n) \int_{\frac{\zeta_\tau}{\sigma}}^\infty y^2 \phi(y) dy$$
$$= (n - p_n) \left(\int_{\frac{\zeta_\tau}{\sigma}}^\infty \phi(y) dy + \frac{\zeta_\tau}{\sigma} \phi\left(\frac{\zeta_\tau}{\sigma}\right) \right) \gtrsim (n - p_n) \zeta_\tau \phi\left(\frac{\zeta_\tau}{\sigma}\right)$$

$$= (n - p_n)\frac{1}{\sqrt{2\pi}}\tau\zeta_\tau. \tag{1.38}$$

By combining the lower bounds (1.37) and (1.38) with the upper bound (1.2), we arrive at (1.7). For the posterior variance, we already have $\sum_{i=p_n+1}^{n}\mathrm{var}(\theta_i \mid Y_i) \asymp (n-p_n)\tau\zeta_\tau$ by (1.24) and (1.26). Expression (1.8) can therefore be proven by showing that there exists a positive constant $c_2(\gamma)$ such that:

$$\mathbb{E}_{\theta=\gamma\zeta_\tau}\mathrm{var}(\theta \mid Y) = \tau^{(1-\gamma)^2}\zeta_\tau^{2\gamma-1}c_2(\gamma)(1+o(1)). \tag{1.39}$$

Proof of (1.36)

The expected value $\mathbb{E}_{\theta=\gamma\zeta_\tau}T_\tau(Y)$ is equal to

$$\frac{1}{\sigma}\left(\int_{-\infty}^{-\frac{\zeta_\tau}{2}} + \int_{-\frac{\zeta_\tau}{2}}^{3\zeta_\tau} + \int_{3\zeta_\tau}^{\infty}\right)(\zeta_\tau + y)\frac{I_{\frac{1}{2}}(\zeta_\tau + y)}{I_{-\frac{1}{2}}(\zeta_\tau + y)}\phi\left(\frac{y + (1-\gamma)\zeta_\tau}{\sigma}\right)dy. \tag{1.40}$$

We shall show that the first and third integrals are negligible, while the second gives the approximation in (1.36). On the domain of the second integral, we have $\zeta_\tau/4 \le \zeta_\tau + y \le 4\zeta_\tau$, so we can apply Lemma 1.12 to see that this integral is asymptotic to

$$\frac{1}{\sigma}\int_{-\frac{\zeta_\tau}{2}}^{3\zeta_\tau}(\zeta_\tau + y)\frac{c_2\tau^2(\zeta_\tau+y)^2 + 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}}{c_1(y+\zeta_\tau)^2 + 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}}\phi\left(\frac{y+(1-\gamma)\zeta_\tau}{\sigma}\right)dy, \tag{1.41}$$

where $c_1 = \int_0^\infty z^{-1/2}(1-z)^{-1}dz$ and $c_2 = \int_0^1 z^{1/2}(1-z)^{-1}dz$. On $[-\zeta_\tau/2, 3\zeta_\tau]$:

$$\begin{aligned} c_2\tau^2(\zeta_\tau+y)^2\phi\left(\frac{y+(1-\gamma)\zeta_\tau}{\sigma}\right) &\le \frac{c_2}{\sqrt{2\pi}}\tau^2(4\zeta_\tau)^3 e^{\frac{(1/2-\gamma)^2\zeta_\tau^2}{2\sigma^2}} \\ &= \frac{64c_2}{\sqrt{2\pi}}\zeta_\tau^3\tau^{2-(1/2-\gamma)^2}, \end{aligned}$$

so (1.41) is asymptotic to:

$$O(\tau) + \frac{2\sigma}{\sqrt{2\pi}}e^{-\frac{(1-\gamma)^2\zeta_\tau^2}{2\sigma^2}}\int_{-\frac{\zeta_\tau}{2}}^{3\zeta_\tau}\frac{(\zeta_\tau+y)e^{\frac{\gamma\zeta_\tau y}{\sigma^2}}}{c_1(y+\zeta_\tau)^2 + 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}}dy.$$

By the substitution $u = \zeta_\tau y - 2\sigma^2\log\zeta_\tau$, the remaining integral is equal to, with $a_\tau = -\frac{\zeta_\tau^2}{2} - 2\sigma^2\log\zeta_\tau$ and $b_\tau = 3\zeta_\tau^2 - 2\sigma^2\log\zeta_\tau$:

$$\begin{aligned} &\frac{2\sigma}{\sqrt{2\pi}}\tau^{(1-\gamma)^2}\frac{1}{\zeta_\tau}\int_{a_\tau}^{b_\tau}\frac{(\zeta_\tau + \zeta_\tau^{-1}(u+2\sigma^2\log\zeta_\tau))e^{\frac{\gamma u}{\sigma^2}}\zeta_\tau^{2\gamma}}{c_1(\zeta_\tau+\zeta_\tau^{-1}(u+2\sigma^2\log\zeta_\tau))^2 + 2\sigma^2 e^{\frac{u}{\sigma^2}}\zeta_\tau^2 e^{\frac{(u+2\sigma^2\log\zeta_\tau)^2}{2\sigma^2\zeta_\tau^2}}}du \\ &\sim \frac{2\sigma}{\sqrt{2\pi}}\tau^{(1-\gamma)^2}\frac{1}{\zeta_\tau}\int_{-\infty}^{\infty}\frac{\zeta_\tau e^{\frac{\gamma u}{\sigma^2}}\zeta_\tau^{2\gamma}}{(c_1 + 2\sigma^2 e^{\frac{u}{\sigma^2}})\zeta_\tau^2}du, \end{aligned}$$

by the dominated convergence theorem. This yields the approximation in (1.36), with $c_1(\gamma) = (2\sigma/\sqrt{2\pi}) \int_{-\infty}^{\infty} e^{\frac{\gamma u}{\sigma^2}}/(c_1 + 2\sigma^2 e^{\frac{u}{\sigma^2}})du$.

For the first integral in (1.40), we use bound 1 from Lemma 1.8, and obtain a bound on its absolute value equal to

$$\frac{1}{\sigma}\int_{-\infty}^{-\frac{\zeta_\tau}{2}} |\zeta_\tau + y|\tau e^{\frac{(\zeta_\tau+y)^2}{2\sigma^2}} \phi\left(\frac{y + (1-\gamma)\zeta_\tau}{\sigma}\right) dy$$
$$= \frac{2}{3\sqrt{2\pi}\sigma}\tau^{(1-\gamma)^2}\int_{-\infty}^{-\frac{\zeta_\tau}{2}} |\zeta_\tau + y| e^{\frac{\gamma\zeta_\tau y}{\sigma^2}} dy \lesssim \tau^{(1-\gamma)^2} e^{-\frac{\gamma\zeta_\tau^2}{2\sigma^2}} = \tau^{(1-\gamma)^2+\gamma}, \quad (1.42)$$

where the last inequality follows by integration by parts. This is of much smaller order than the second integral from (1.40). In the third integral of (1.40), we bound $I_{\frac{1}{2}}(\zeta_\tau + y)/I_{-\frac{1}{2}}(\zeta_\tau + y)$ by 1, giving the upper bound

$$\frac{1}{\sigma}\int_{3\zeta_\tau}^{\infty} (\zeta_\tau + y)\phi\left(\frac{y + (1-\gamma)\zeta_\tau}{\sigma}\right) dy \lesssim \phi\left(\frac{3\zeta_\tau + (1-\gamma)\zeta_\tau}{\sigma}\right) = \frac{1}{\sqrt{2\pi}}\tau^{4-\gamma},$$

by Mills' ratio. This is also of much smaller order than the second integral from (1.40), thus concluding the proof of (1.36).

Proof of (1.39)

By expanding the term $(1-z)^2$ in the numerator of the final term of (1.20), the posterior variance can be seen to be equal to:

$$\mathrm{var}(\theta \mid y) = \sigma^2 \frac{I_{\frac{1}{2}}(y)}{I_{-\frac{1}{2}}(y)} + y^2\left[\frac{I_{\frac{3}{2}}(y)}{I_{-\frac{1}{2}}(y)} - \left(\frac{I_{\frac{1}{2}}(y)}{I_{-\frac{1}{2}}(y)}\right)^2\right]. \quad (1.43)$$

Because $I_{\frac{1}{2}}(y)/I_{-\frac{1}{2}}(y)$ can be interpreted as the mean of the density proportional to $z \to z^{-1/2}e^{y^2 z/(2\sigma^2)}/(\tau^2 + (1-\tau^2)z)$, and $I_{\frac{3}{2}}(y)/I_{-\frac{1}{2}}(y)$ as the second moment, it follows that the term in square brackets in (1.43) is nonnegative. By (1.43), we write:

$$\mathbb{E}_{\theta=\gamma\zeta_\tau}\mathrm{var}(\theta \mid Y) = \sigma\int \frac{I_{\frac{1}{2}}(\zeta_\tau + y)}{I_{-\frac{1}{2}}(\zeta_\tau + y)}\phi\left(\frac{y + (1-\gamma)\zeta_\tau}{\sigma}\right) dy$$
$$+ \frac{1}{\sigma}\left(\int_{-\infty}^{-\frac{\zeta_\tau}{2}} + \int_{-\frac{\zeta_\tau}{2}}^{3\zeta_\tau} + \int_{3\zeta_\tau}^{\infty}\right)(\zeta_\tau + y)^2$$
$$\cdot\left[\frac{I_{\frac{3}{2}}(\zeta_\tau + y)}{I_{-\frac{1}{2}}(\zeta_\tau + y)} - \left(\frac{I_{\frac{1}{2}}(\zeta_\tau + y)}{I_{-\frac{1}{2}}(\zeta_\tau + y)}\right)^2\right]\phi\left(\frac{y + (1-\gamma)\zeta_\tau}{\sigma}\right) dy. \quad (1.44)$$

The first term in (1.44) is as (1.40), except without the factor $(\zeta_\tau + y)$. Following the same steps as the proof of (1.36), we see that it is smaller than a multiple of ζ_τ^{-1} times the bound on (1.40), so it is of the order $\zeta_\tau^{2\gamma-3}\tau^{(1-\gamma)^2}$. The first and third integrals of the second term of (1.44) are also negligible. For the first, we use that the expression in square brackets is nonnegative and bounded above by $I_{\frac{3}{2}}(y)/I_{-\frac{1}{2}}(y)$, which in turn is bounded above by

$I_{\frac{1}{2}}(y)/I_{-\frac{1}{2}}(y)$. We bound as in (1.42), with the difference that the leading factor is $(\zeta_\tau + y)^2$ instead of $(\zeta_\tau + y)$. This leads to the order $\zeta_\tau \tau^{(1-\gamma)^2+\gamma}$, much smaller than the claimed rate. For the third integral, we can bound the term in square brackets by 1 and use Mills' ratio to see that it is of the order $\zeta_\tau \tau^{(4-\gamma)^2}$.

We are left with the middle integral of the second term of (1.44). On the domain of this integral, by Lemma 1.12:

$$\frac{I_{\frac{3}{2}}(\zeta_\tau + y)}{I_{-\frac{1}{2}}(\zeta_\tau + y)} = \frac{c_3 \tau^4 (\zeta_\tau + y)^2 + 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}}{c_1(\zeta_\tau + y)^2 + 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}}(1 + o(1)),$$

where $c_3 = \int_0^1 z^{3/2}(1+z)^{-1}dz$, and c_1 is as in (1.41). We see that $I_{\frac{3}{2}}(y)/I_{-\frac{1}{2}}(y)$ and $I_{\frac{1}{2}}(y)/I_{-\frac{1}{2}}(y)$ are asymptotic to the same function on this domain. Since $A/(A+B) - A^2/(A+B)^2 = AB/(A+B)^2$, it follows that up to $O(\tau)$, the middle integral is asymptotic to

$$\begin{aligned}
&\frac{1}{\sigma}\int_{-\frac{\zeta_\tau}{2}}^{3\zeta_\tau} (\zeta_\tau + y)^2 \frac{c_1(\zeta_\tau + y)^2 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}}{\left(c_1(\zeta_\tau + y)^2 + 2\sigma^2 e^{\frac{y^2+2y\zeta_\tau}{2\sigma^2}}\right)^2} \phi\left(\frac{y + (1-\gamma)\zeta_\tau}{\sigma}\right) dy \\
&\qquad = \frac{2\sigma c_1}{\sqrt{2\pi}} \tau^{(1-\gamma)^2} \int_{-\frac{\zeta_\tau}{2}}^{3\zeta_\tau} \frac{(\zeta_\tau + y)^4 e^{\frac{\gamma\zeta_\tau y}{\sigma^2}}}{\left(c_1(\zeta_\tau + y)^2 + 2\sigma^2 e^{\frac{y^2+2\zeta_\tau y}{2\sigma^2}}\right)^2} dy.
\end{aligned}$$

We substitute $u = \zeta_\tau y - 2\sigma^2 \log \zeta_\tau$ to reduce this to

$$\begin{aligned}
&\frac{2\sigma c_1}{\sqrt{2\pi}} \tau^{(1-\gamma)^2} \frac{1}{\zeta_\tau} \int_{-\frac{\zeta_\tau^2}{2} - 2\sigma^2 \log \zeta_\tau}^{3\zeta_\tau^2 - 2\sigma^2 \log \zeta_\tau} \frac{(\zeta_\tau + \zeta_\tau^{-1}(u + 2\sigma^2 \log \zeta_\tau))^4 e^{\frac{\gamma u}{\sigma^2}} \zeta_\tau^{2\gamma}}{\left(c_1(\zeta_\tau + \zeta_\tau^{-1}(u + 2\sigma^2 \log \zeta_\tau))^2 + 2\sigma^2 e^{\frac{u}{\sigma^2}} \zeta_\tau^2\right)^2} du \\
&\qquad \sim \frac{2\sigma c_1}{\sqrt{2\pi}} \tau^{(1-\gamma)^2} \frac{1}{\zeta_\tau} \int_{-\infty}^{\infty} \frac{\zeta_\tau^4 e^{\frac{\gamma u}{\sigma^2}} \zeta_\tau^{2\gamma}}{\left(c_1 \zeta_\tau^2 + 2\sigma^2 \zeta_\tau^2 e^{\frac{u}{\sigma^2}}\right)^2} du.
\end{aligned}$$

This is asymptotic to expression (1.39), with $c_2(\gamma) = (2\sigma c_1/\sqrt{2\pi}) \int_{-\infty}^{\infty} e^{\frac{\gamma u}{\sigma^2}}/(c_1 + 2\sigma^2 e^{\frac{u}{\sigma^2}})^2 du$. □

Proof of Theorem 1.6

Proof. Suppose that $Y \sim \mathcal{N}(\theta, \sigma^2 I_n)$, $\theta \in \ell_0[p_n]$. We adapt the approach in Paragraph 6.2 in (Johnstone and Silverman, 2004). We first derive the following inequality for events A such that $\widehat{\tau} > \tau$ holds with probability one on A:

$$\begin{aligned}
\mathbb{E}_\theta (T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_A &\le 2\mathbb{E}_\theta (T_{\widehat{\tau}}(Y_i) - Y_i)^2 \mathbf{1}_A + 2\mathbb{E}_\theta (Y_i - \theta_i)^2 \mathbf{1}_A \\
&\lesssim 2\mathbb{E}_\theta \zeta_{\widehat{\tau}}^2 \mathbf{1}_A + 2\sigma^2 \mathbb{E}_\theta Z^2 \mathbf{1}_A \qquad (1.45)
\end{aligned}$$

where (1.17) was used in the second line, and Z follows a standard normal distribution. If A is such that $\widehat{\tau} > \tau$ holds with probability one on A, we can use the inequality $\zeta_{\widehat{\tau}} < \zeta_\tau$ if $\widehat{\tau} > \tau$ to find:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_A \lesssim 2\zeta_\tau^2 \mathbb{P}_\theta(A) + 2\sigma^2 \mathbb{E}_\theta Z^2 \mathbf{1}_A, \tag{1.46}$$

We now consider the nonzero and zero parameters separately. For both cases, we split up the expected ℓ_2 loss as follows:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 = \mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} > c\tau\}} + \mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} \le c\tau\}},$$

and then bound each of terms on the right hand side. For the nonzero means, we take $c = 1$, while for the zero means, we consider $c \ge 1$. Note that for $\zeta_{\widehat{\tau}}$ to be well-defined, we need $\widehat{\tau} \le 1$ and consequently, when we consider $\widehat{\tau} > c\tau$, we must have $c\tau < 1$.

Nonzero means

By (1.46), we find:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} > \tau\}} \lesssim 2\zeta_\tau^2 + 2\sigma^2. \tag{1.47}$$

If $\widehat{\tau} \le \tau$, the inequality $\zeta_{\widehat{\tau}}^2 \le \zeta_\tau^2$ needed for (1.46) does not hold. For this case, we assume that $\widehat{\tau} \ge g(n,p_n)$ with probability one, for some function $g(n,p_n)$, corresponding to $\zeta_{\widehat{\tau}} \le \sqrt{-2\sigma^2 \log g(n,p_n)}$. Then we find by (1.45):

$$\begin{aligned}\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} \le \tau\}} &\lesssim 2\mathbb{E}_\theta \zeta_{\widehat{\tau}}^2 \mathbf{1}_{\{\widehat{\tau} \le \tau\}} + 2\sigma^2 \\ &\le -4\sigma^2 \log(g(n,p_n)) \mathbb{P}_\theta(\widehat{\tau} \le \tau) + 2\sigma^2. \end{aligned} \tag{1.48}$$

By (1.47) and (1.48), we have for $\theta_i \neq 0$:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \lesssim 1 + \zeta_\tau^2 - \log(g(n,p_n)) \mathbb{P}_\theta(\widehat{\tau} \le \tau). \tag{1.49}$$

Zero means

We first establish an inequality for $\mathbb{E}_\theta[Z^2 \mathbf{1}_A]$, where A is an event and Z a standard normal random variable. By Young's inequality, we have for any positive x and y:

$$xy \le \int_0^x (e^s - 1) ds + \int_0^y \log(s+1) ds = e^x - x - 1 + (y+1)\log(y+1) - y.$$

By this inequality combined with the inequality $\log(y+1) < y$, we have:

$$\mathbb{E}_\theta Z^2 \mathbf{1}_A \le cd\mathbb{E}_\theta \left[e^{\frac{Z^2}{c}} - \frac{Z^2}{c} - 1 \right] + cd\mathbb{P}_\theta(A) \left(\frac{1}{d} \log\left(\frac{1}{d} + 1 \right) - \frac{1}{d} \right).$$

With $c = 3$ and $d = \mathbb{P}_\theta(A)$, we find:

$$\begin{aligned}\mathbb{E}_\theta Z^2 \mathbf{1}_A &\le (3\sqrt{3} - 4)\mathbb{P}_\theta(A) + 3\mathbb{P}_\theta(A) \log\left(1 + \frac{1}{\mathbb{P}_\theta(A)} \right) \\ &< 5\mathbb{P}_\theta(A) \log\left(1 + \frac{1}{\mathbb{P}_\theta(A)} \right). \end{aligned} \tag{1.50}$$

By (1.46) and (1.50), we get for any $c \geq 1$ such that $c\tau < 1$:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} > c\tau\}} \lesssim 2\zeta_\tau^2 \mathbb{P}_\theta(\widehat{\tau} > c\tau) + 10\sigma^2 \mathbb{P}_\theta(\widehat{\tau} > c\tau) \log\left(1 + \frac{1}{\mathbb{P}_\theta(\widehat{\tau} > c\tau)}\right). \tag{1.51}$$

Now suppose $\widehat{\tau} \leq c\tau$ for some $c \geq 1$ such that $c\tau < 1$. First note that $|T_\tau(y)|$ increases monotonically in τ, as is clear from

$$T_\tau(y_i) = \mathbb{E}[(1-\kappa_i)y_i \mid y_i, \tau] = \mathbb{E}\left[\frac{\tau^2\lambda_i^2}{1+\tau^2\lambda_i^2} y_i \,\middle|\, y_i, \tau\right].$$

Because $\operatorname{sign}(T_{\widehat{\tau}}(y_i)) = \operatorname{sign}(T_{c\tau}(y_i))$ and $0 \leq |T_{\widehat{\tau}}(y_i)| \leq |T_{c\tau}(y_i)|$, we have:

$$(T_{\widehat{\tau}}(y_i) - \theta_i)^2 \leq \max\{\theta_i^2, (T_{c\tau}(y_i) - \theta_i)^2\} \leq \theta_i^2 + (T_{c\tau}(y_i) - \theta_i)^2.$$

Hence:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} \leq c\tau\}} \leq \theta_i^2 + \mathbb{E}_\theta(T_{c\tau}(Y_i) - \theta_i)^2.$$

And thus, by (1.18), we have for $\theta_i = 0$:

$$\mathbb{E}_\theta(T_{\widehat{\tau}}(Y_i) - \theta_i)^2 \mathbf{1}_{\{\widehat{\tau} \leq c\tau\}} \lesssim \zeta_{c\tau} c\tau \lesssim \zeta_\tau \tau. \tag{1.52}$$

Combining (1.51) and (1.52), we find:

$$\mathbb{E}_\theta T_{\widehat{\tau}}(Y_i)^2 \lesssim \zeta_\tau \tau + \zeta_\tau^2 \mathbb{P}_\theta(\widehat{\tau} > c\tau) + \mathbb{P}_\theta(\widehat{\tau} > c\tau) \log\left(1 + \frac{1}{\mathbb{P}_\theta(\widehat{\tau} > c\tau)}\right). \tag{1.53}$$

Conclusion

We can now bound the expected ℓ_2 loss. We assume that $\theta_i \neq 0$ for $i = 1, \ldots, \tilde{p}_n$ and $\theta_i = 0$ for $i = \tilde{p}_n + 1, \ldots, n$, where $\tilde{p}_n \leq p_n$. By combining (1.49) and (1.53), we find:

$$\mathbb{E}_\theta \|T_{\widehat{\tau}}(Y) - \theta\|^2 \lesssim \tilde{p}_n \left(1 + \zeta_\tau^2 - \log(g(n,p_n)) \mathbb{P}_\theta(\widehat{\tau} \leq \tau)\right) + (n - \tilde{p}_n)\zeta_\tau \tau + (n - \tilde{p}_n)\mathbb{P}_\theta(\widehat{\tau} > c\tau)\left(\zeta_\tau^2 + \log\left(1 + \frac{1}{\mathbb{P}_\theta(\widehat{\tau} > c\tau)}\right)\right). \tag{1.54}$$

The function $x \log(1 + \frac{1}{x})$ is monotonically increasing in x for $x \in [0,1]$. Hence, with the choice $\tau = \frac{p_n}{n}$ or $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$, the conditions stated in the theorem are sufficient for (1.54) to be bounded by the minimax squared error rate in the worst case.

If an estimator $\widehat{\tau}$ satisfies only the first condition, then $\sup\{\frac{1}{n}, \widehat{\tau}\}$ satisfies the second condition with $-\log g(n,p_n) = \log n$. By the assumption $p_n \to \infty$, we have $\mathbb{P}_\theta(\sup\{\frac{1}{n}, \widehat{\tau}\} > c\frac{p_n}{n}) \leq \mathbb{P}_\theta(\widehat{\tau} > c\frac{p_n}{n})$. Plugging this into inequality (1.54) yields an ℓ_2 risk of at most order $p_n \log n$. □

Lemma 1.13. *Suppose $Y_i \sim \mathcal{N}(\theta_i, \sigma^2), i = 1, \ldots, p_n$ and $Y_i \sim \mathcal{N}(0, \sigma^2), i = p_n + 1, \ldots, n$ and define*

$$\widehat{\tau} = \frac{\#\{|y_i| \geq \sqrt{c_1 \sigma^2 \log n}, i = 1, \ldots, n\}}{c_2 n}$$

for some $c_2 > 1$. Then $\mathbb{P}_\theta(\widehat{\tau} > \tau) \lesssim \frac{p_n}{n}$ as $p_n, n \to \infty$, $p_n = o(n)$ if $c_1 > 2$, or $c_1 = 2$ and $p_n \lesssim \log n$ for $\tau = \frac{p_n}{n}$ or $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$.

Proof. We only need to consider $\mathbb{P}_\theta(\widehat{\tau} > \frac{p_n}{n})$, as we assume $p_n = o(n)$ and thus, for large n, $\mathbb{P}_\theta(\widehat{\tau} > \frac{p_n}{n}\sqrt{\log(n/p_n)}) \leq \mathbb{P}_\theta(\widehat{\tau} > \frac{p_n}{n})$. Define $A_i = \{|y_i| \geq \sqrt{c_1\sigma^2 \log n}\}, i = 1,\dots,n$. For $i = p_n + 1,\dots,n$, $\mathbf{1}_{A_i}$ follows a Bernoulli distribution with parameter $q_n = 2\Phi^c(\sqrt{c_1 \log n})$, which by Mills' ratio can be bounded from above by $\sqrt{\frac{2}{c_1\pi}}(\log n)^{-\frac{1}{2}} n^{-\frac{c_1}{2}}$. For $X \sim \mathrm{Bin}(n,p)$, we have the bound $\mathbb{P}(X \geq k) \leq (\frac{enp}{k})^k$ as a consequence of Theorem 1 in (Chernoff, 1952). Hence:

$$\begin{aligned}\mathbb{P}_\theta\left(\widehat{\tau} > \frac{p_n}{n}\right) &\leq \mathbb{P}_\theta\left(\sum_{i=p_n+1}^{n} \mathbf{1}_{A_i} > (c_2-1)p_n\right) \leq \left(\frac{e(n-p_n)q_n}{(c_2-1)p_n+1}\right)^{(c_2-1)p_n+1} \\ &\leq \left(\sqrt{\frac{2e^2}{c_1\pi}}\frac{1}{(c_2-1)p_n+1}\frac{1}{\sqrt{\log n}}n^{1-\frac{c_1}{2}}\right)^{(c_2-1)p_n+1}. \end{aligned} \tag{1.55}$$

The inequality $\mathbb{P}_\theta(\widehat{\tau} > \frac{p_n}{n}) \lesssim \frac{p_n}{n}$ holds if $-\log \mathbb{P}_\theta(\widehat{\tau} > \frac{p_n}{n}) \geq \log \frac{n}{p_n} + c$ holds for some positive constant c. The negative logarithm of bound (1.55) is:

$$((c_2-1)p_n+1)\left(\frac{1}{2}\log\frac{c_1\pi}{2e^2} + \log((c_2-1)p_n+1) + \frac{1}{2}\log\log n + \left(\frac{c_1}{2}-1\right)\log n\right).$$

For $c_1 = 2$, this quantity will exceed $\log \frac{n}{p_n}$ if $p_n \gtrsim \log n$. If $c_1 > 2$, we require $((c_2-1)p_n + 1)(\frac{c_1}{2} - 1) \geq 1$, which is certainly satisfied if $p_n \to \infty$. □

2

Conditions for posterior concentration for scale mixtures of normals

Abstract

The first Bayesian results for the sparse normal means problem were proven for spike-and-slab priors. However, these priors are less convenient from a computational point of view. In the meanwhile, a large number of continuous shrinkage priors has been proposed. Many of these shrinkage priors can be written as a scale mixture of normals, which makes them particularly easy to implement. We propose general conditions on the prior on the local variance in scale mixtures of normals, such that posterior contraction at the minimax rate is assured. The conditions require tails at least as heavy as Laplace, but not too heavy, and a large amount of mass around zero relative to the tails, more so as the sparsity increases. These conditions give some general guidelines for choosing a shrinkage prior for estimation under a nearly black sparsity assumption. We verify these conditions for the class of priors considered in Ghosh and Chakrabarti (2015), which includes the horseshoe and the normal-exponential gamma priors, and for the horseshoe+, the inverse-Gaussian prior, the normal-gamma prior, and the spike-and-slab Lasso, and thus extend the number of shrinkage priors which are known to lead to posterior contraction at the minimax estimation rate.

This chapter has appeared as S.L. van der Pas, J.-B. Salomond and J. Schmidt-Hieber (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics* **10**, 976–1000. Research supported by NWO VICI project 'Safe Statistics'.

2.1 Introduction

In the sparse normal means problem, we wish to estimate a sparse vector θ based on a vector $X^n \in \mathbb{R}^n$, $X^n = (X_1, \ldots, X_n)$, generated according to the model

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the ε_i are independent standard normal variables. The vector of interest θ is sparse in the *nearly black* sense, that is, most of the parameters are zero. We wish to separate the signals (nonzero means) from the noise (zero means). Applications of this model include image reconstruction and nonparametric function estimation using wavelets (Johnstone and Silverman, 2004).

The model is an important test case for the behaviour of sparsity methods, and has been well-studied. A great variety of frequentist and Bayesian estimators has been proposed, and the popular Lasso (Tibshirani, 1996) is included in both categories. It is but one example of many approaches towards recovering θ; restricting ourselves to Bayesian methods, other approaches include shrinkage priors such as the spike-and-slab type priors studied by Castillo and Van der Vaart (2012); Johnstone and Silverman (2004) and Castillo et al. (2015), the normal-gamma prior (Griffin and Brown, 2010), non-local priors (Johnson and Rossell, 2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2014), the horseshoe (Carvalho et al., 2010), the horseshoe+ (Bhadra et al., 2015) and the spike-and-slab Lasso (Ročková, 2015).

Our goal is twofold: *recovery* of the underlying mean vector, and *uncertainty quantification*. The benchmark for the former is estimation at the minimax rate. In a Bayesian setting, the typical choice for the estimator is some measure of center of the posterior distribution, such as the posterior mean, mode or median. For the purpose of uncertainty quantification, the natural object to use is a credible set. In order to obtain credible sets that are narrow enough to be informative, yet not so narrow that they neglect to cover the truth, the posterior distribution needs to contract to its center at the same rate at which the estimator approaches the truth.

For recovery, spike-and-slab type priors give optimal results (Castillo et al. (2015); Castillo and Van der Vaart (2012); Johnstone and Silverman (2004)). These priors assign independently to each component a mixture of a point mass at zero and a continuous prior. Due to the point mass, spike-and-slab priors shrink small coefficients to zero. The advantage is that the full posterior has optimal model selection properties but this comes at the price of, in general, too narrow credible sets. Another drawback of spike-and-slab methods is that they are computationally expensive although the complexity is much better than what has been previously believed (Yang et al. (2015)).

Thus, we might ask whether there are priors which are smoother and shrink less than the spike-and-slab but still recover the signal with a (nearly) optimal rate. A naive choice would be to consider the Laplace prior $\propto e^{-\lambda\|\theta\|_1}$ with $\|\theta\|_1 = \sum_{i=1}^n |\theta_i|$, since in this case the maximum a posteriori (MAP) estimator coincides with the Lasso, which is known to achieve the optimal rates for sparse signals. In Castillo et al. (2015), Section 3, it was shown that although the MAP-estimator has good properties, the full posterior spreads a non-negligible amount of mass over large neighborhoods of the truth leading to recovery

rates that are sub-optimal by a polynomial factor in n. This example shows that if the prior does not shrink enough, we loose the recovery property of the posterior.

Recently, shrinkage priors were found that are smoother than the spike-and-slab but still lead to (near) minimax recovery rates. Up to now, optimal recovery rates have been established for the horseshoe prior (Van der Pas et al., 2014), horseshoe-type priors with slowly varying functions (Ghosh and Chakrabarti, 2015), the empirical Bayes procedure of Martin and Walker (2014), the spike-and-slab Lasso (Ročková, 2015), and the Dirichlet-Laplace prior, although the latter result only holds under a restriction on the signal size (Bhattacharya et al., 2014). Finding smooth shrinkage priors with theoretical guarantees remains an active area of research.

The question arises which features of the prior lead to posterior convergence at the minimax estimation rate. Qualitative discussion on this point is provided by Carvalho et al. (2010). Intuitively, a prior should place a large amount of mass near zero to account for the zero means, and have heavy tails to counteract the shrinkage effect for the nonzero means. In the present article, we make an attempt to quantify the relevant properties of a prior, by providing general conditions ensuring posterior concentration at the minimax rate, and showing that a large number of priors (including the ones listed above) meets these conditions.

We study scale mixtures of normals, as many shrinkage priors proposed in the literature are contained in this class and provide general conditions on the prior on the local variance such that posterior concentration at the minimax estimation rate is guaranteed. These conditions are general enough to recover the already known results for the horseshoe prior, the horseshoe-type priors with slowly varying functions and the spike-and-slab Lasso, and to demonstrate that the horsehoe+ (Bhadra et al., 2015), inverse-Gaussian prior (Caron and Doucet, 2008) and the normal-gamma prior (Caron and Doucet, 2008; Griffin and Brown, 2010) lead to posterior concentration at the correct rate as well. Our conditions in essence mean that a sparsity prior should have tails that are at least as heavy as Laplace, but not too heavy, and there should be a sizable amount of mass close to zero relative to the tails, especially when the underlying vector is very sparse.

This paper is organized as follows. We state our main result, providing conditions on sparsity priors such that the posterior contracts at the minimax rate in Section 2.2. We then show, in Section 2.3, that these conditions hold for the class of priors of Ghosh and Chakrabarti (2015), as well as for the horseshoe+, the inverse-Gaussian prior, the normal-gamma prior, and the spike-and-slab Lasso. A simulation study is performed in Section 2.4, and we conclude with a Discussion. All proofs are given in Appendix 2.6.

Notation. Denote the class of nearly black vectors by $\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \mathbf{1}\{\theta_i \neq 0\} \leq p_n\}$. The minimum $\min\{a, b\}$ is given by $a \wedge b$. The standard normal density is denoted by ϕ, its cdf by Φ, and we set $\Phi^c(x) = 1 - \Phi(x)$. The norm $\|\cdot\|$ is the ℓ_2-norm.

2.2 Main results

Each coefficient θ_i receives a scale mixture of normals as a prior:

$$\theta_i \mid \sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2), \qquad \sigma_i^2 \sim \pi(\sigma_i^2), \quad i = 1, \ldots, n, \tag{2.1}$$

where $\pi : [0,\infty) \to [0,\infty)$ is a density on the positive reals. While π might depend on further hyperparameters, no additional priors are placed on such parameters, rendering the coefficients independent *a posteriori*. The goal is to obtain conditions on π such that posterior concentration at the minimax estimation rate is guaranteed.

We use the coordinatewise posterior mean to recover the underlying mean vector. By Tweedie's formula (Robbins, 1956), the posterior mean for θ_i given an observation x_i is equal to $x_i + \frac{d}{dx}\log p(x_i)$, where $p(x_i)$ is the marginal distribution of x_i. The posterior mean for parameter θ_i is thus given by $\widehat{\theta_i} = X_i m_{X_i}$, where $m_x : \mathbb{R} \to [0,1]$ is

$$m_x := \frac{\int_0^1 z(1-z)^{-3/2} e^{\frac{x^2}{2}z} \pi\left(\frac{z}{1-z}\right) dz}{\int_0^1 (1-z)^{-3/2} e^{\frac{x^2}{2}z} \pi\left(\frac{z}{1-z}\right) dz} = \frac{\int_0^\infty u(1+u)^{-3/2} e^{\frac{x^2 u}{2+2u}} \pi(u) du}{\int_0^\infty (1+u)^{-1/2} e^{\frac{x^2 u}{2+2u}} \pi(u) du}. \tag{2.2}$$

We denote the estimate of the full vector θ by $\widehat{\theta} = (\widehat{\theta_1}, \ldots, \widehat{\theta_n}) = (X_1 m_{X_1}, \ldots, X_n m_{X_n})$. An advantage of scale mixtures of normals as shrinkage priors over spike-and-slab-type priors, is that the posterior mean can be represented as the observation multiplied by (2.2). The ratio (2.2) can be computed via integral approximation methods such as a quadrature routine. See Polson and Scott (2012*a*), Polson and Scott (2012*b*) and Van der Pas et al. (2014) for more discussion on this point in the context of the horseshoe.

Our main theorem, Theorem 2.1, provides three conditions on π under which a prior of the form (2.1) leads to an upper bound on the posterior contraction rate of the order of the minimax rate. We first state and discuss the conditions. In addition, we present stronger conditions that are easier to verify. Condition 1 is required for our bounds on the posterior mean and variance for the nonzero means. The remaining two are used for the bounds for the zero means.

The first condition involves a class of regularly varying functions. Recall that a function ℓ is called *regular varying (at infinity)* if for any $a > 0$, the ratio $\ell(au)/\ell(u)$ converges to the same non-zero limit as $u \to \infty$. For our estimates, we need a slightly different notion, that will be introduced next. We say that a function L is *uniformly regular varying,* if there exist constants $R, u_0 \geq 1$, such that

$$\frac{1}{R} \leq \frac{L(au)}{L(u)} \leq R, \quad \text{for all } a \in [1,2], \text{ and all } u \geq u_0. \tag{2.3}$$

In particular, $L(u) = u^b$, and $L(u) = \log^b(u)$ with $b \in \mathbb{R}$ are uniformly regular varying (take for example $R = 2^{|b|}$ and $u_0 = 2$). An example of a function that is not uniformly regular varying is $L(u) = e^u$. From the definition, we can easily deduce the following properties of functions that are uniformly regular varying. Firstly, $u \mapsto L(u)$ is on $[u_0,\infty)$ either everywhere positive or everywhere negative. If L is uniformly regular varying then so is $u \mapsto 1/L(u)$ and if L_1 and L_2 are uniformly regular varying, then so is their product $L_1 L_2$.

We are now ready to present Condition 1, and the stronger Condition 1', which implies Condition 1, as shown in Lemma 2.3.

Condition 1. For some $b \geq 0$, we can write $u \mapsto \pi(u) = L_n(u) e^{-bu}$, where L_n is a function that satisfies (2.3) for some $R, u_0 \geq 1$ which do not depend on n. Suppose further that there

are constants $C', b' > 0, K \geq 0$, and $u_* \geq 1$, such that

$$C'\pi(u) \geq \left(\frac{p_n}{n}\right)^K e^{-b'u} \quad \text{for all } u \geq u_*. \tag{2.4}$$

Condition 1'. Consider a global-local scale mixture of normals:

$$\theta_i \mid \sigma_i^2, \tau^2 \sim \mathcal{N}(0, \sigma_i^2\tau^2), \quad \sigma_i^2 \sim \widetilde{\pi}(\sigma_i^2), \quad i = 1, \ldots, n. \tag{2.5}$$

Assume that $\widetilde{\pi}$ is a uniformly regular varying function which does not depend on n, and $\tau = (p_n/n)^\alpha$ for $\alpha \geq 0$.

Condition 1 assures that the posterior recovers nonzero means with the optimal rate. Thus, the condition can be seen as a sufficient condition on the tail behavior of the density π for ℓ^2-recovery. The tail may decay exponentially fast, which is consistent with the conditions found on the 'slab' in the spike-and-slab priors discussed by Castillo and Van der Vaart (2012). In general, π will depend on n through a hyperparameter. Condition 1 requires that the n dependence behaves roughly as a power of p_n/n.

In the important special case where each θ_i is drawn independently from a global-local scale mixture, Condition 1 is satisfied whenever the density on the local variance is uniformly regular varying, as stated in Condition 1'. Below, we give the conditions on π that guarantee posterior shrinkage at the minimax rate for the zero coefficients. The first condition ensures that the prior π puts some finite mass on values between $[0, 1]$.

Condition 2. Suppose that there is a constant $c > 0$ such that $\int_0^1 \pi(u)du \geq c$.

We turn to Condition 3 which describes the decay of π away from a neighborhood of zero. To state the condition it will be convenient to write

$$s_n := \frac{p_n}{n}\log(n/p_n). \tag{2.6}$$

Condition 3. Let $b_n = \sqrt{\log(n/p_n)}$ and assume that there is a constant C, such that

$$\int_{s_n}^{\infty} \left(u \wedge \frac{b_n^3}{\sqrt{u}}\right)\pi(u)du + b_n \int_1^{b_n^2} \frac{\pi(u)}{\sqrt{u}}du \leq Cs_n.$$

In order to allow for many possible choices of π, the tail condition involves several terms. Observe that $u \wedge b_n^3/\sqrt{u} = u$ if and only if $u \leq b_n^2$ and therefore the first integral in Condition 3 can also be written as $\int_{s_n}^{b_n^2} u\pi(u)du + b_n^3 \int_{b_n^2}^{\infty} u^{-1/2}\pi(u)du$. It is surprising that some control of $\pi(u)$ on the interval $[s_n, 1]$ is needed. But this turns out to be sharp. Theorem 2.2 proves that if we would relax the condition to $\int_{s_n}^1 u\pi(u)du \lesssim t_n$ for an arbitrary rate $t_n \gg s_n$, then there is a prior that satisfies all the other conditions needed for the zero coefficients, but which does not lead to concentration at the minimax rate.

Below we state two stronger conditions, each of which obviously implies Condition 2 and Condition 3 for sparse signals, that is, $p_n = o(n)$.

Condition A. Assume that there is a constant C, such that

$$\pi(u) \leq \frac{C}{u^{3/2}} \frac{p_n}{n} \sqrt{\log(n/p_n)}, \quad \text{for all } u \geq s_n.$$

Condition B. Assume that there is a constant C, such that

$$\int_{s_n}^{\infty} \pi(u) du \leq \frac{C p_n}{n}.$$

In this case, even a stronger version of Condition 2 holds in the sense that nearly all mass is concentrated in the shrinking interval $[0, s_n]$. Notice that Condition 3 does not imply Condition 2 in general. If, for example, the density π has support on $[n^2, 2n^2]$, then, Condition 3 holds but Condition 2 does not. Condition 1 and Condition 3 depend on the relative sparsity p_n/n. Indeed, Condition 1 becomes weaker if the signal is more sparse and at the same time Condition 3 becomes stronger. This matches intuition, as the prior should shrink more in this case and thus the assumptions that are responsible for the shrinkage effect should become stronger.

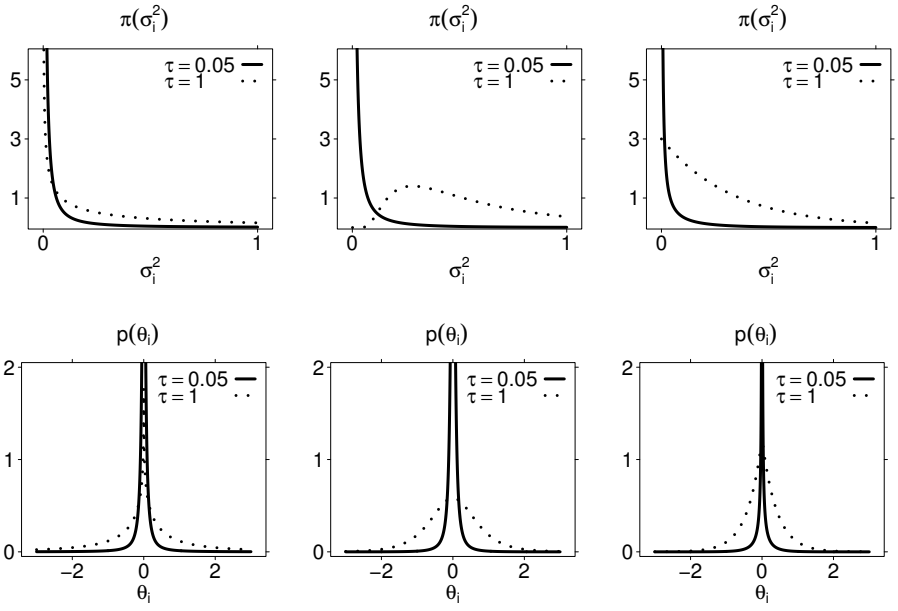

Figure 2.1: Plots of priors on the local variance (first row) and the corresponding parameters (second row). From left to right: horseshoe, Inverse-Gaussian with $a = 1/2, b = 1$, and normal gamma with $\beta = 3$. The parameter τ, which in practice should be of the order p_n/n, is taken equal to 1 (dashed line) and 0.05 (solid line).

Figure 2.1 presents plots of the priors π on the local variance, and the corresponding priors on the parameters θ_i, for three priors for which the three conditions are verified in

Section 2.3: the horseshoe, inverse-Gaussian, and normal-gamma. The parameter τ, in the notation of Section 2.3, should be thought of as the sparsity level p_n/n. Figure 2.1 shows that the priors start to resemble each other when τ is decreased. If the setting is more sparse, corresponding to more zero means, the mass of the prior π on σ_i^2 concentrates around zero, leading to a higher peak at zero in the prior density on θ_i.

We now present our main result. The minimax estimation risk for this problem, under ℓ_2 risk, is given by $2p_n \log(n/p_n)$ Donoho et al. (1992). We write $\theta_0 = (\theta_{0i})_{i=1,\dots,n}$ and consider posterior concentration of the zero and non-zero coefficients separately. Asymptotics always refers to $n \to \infty$.

Theorem 2.1. *Work under model $X^n \sim \mathcal{N}(\theta_0, I_n)$ and assume that the prior is of the form (2.1). Suppose further that $p_n = o(n)$ and let M_n be an arbitrary positive sequence tending to $+\infty$. Let $\widehat{\theta} = (\widehat{\theta_1}, \dots, \widehat{\theta_n})$ be the posterior mean. Under Condition 1,*

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi\Big(\theta \,:\, \sum_{i:\theta_{0i}\neq 0} (\theta_i - \theta_{0i})^2 > M_n p_n \log(n/p_n) \,\Big|\, X^n\Big) \to 0$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \sum_{i:\theta_{0i}\neq 0} (\widehat{\theta_i} - \theta_{0i})^2 \lesssim p_n \log(n/p_n).$$

Under Condition 2 and Condition 3 (or either Condition A or B),

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi\Big(\theta \,:\, \sum_{i:\theta_{0i}=0} \theta_i^2 > M_n p_n \log(n/p_n) \,\Big|\, X^n\Big) \to 0$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \sum_{i:\theta_{0i}=0} \widehat{\theta_i^2} \lesssim p_n \log(n/p_n).$$

Thus, under Conditions 1-3 (or Condition 1 with either Condition A or B),

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi\Big(\theta \,:\, \|\theta - \theta_0\|^2 > M_n p_n \log(n/p_n) \,\Big|\, X^n\Big) \to 0$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Big\|\widehat{\theta} - \theta_0\Big\|_2^2 \lesssim p_n \log(n/p_n).$$

The statement is split into zero and non-zero coefficients of θ_0 in order to make the dependence on the conditions explicit. Indeed, posterior concentration of the non-zero coefficients follows from Condition 1 and posterior concentration for the zero-coefficients is a consequence of Conditions 2 and 3. In order to obtain posterior contraction, we need that $M_n \to \infty$. This is due to the use of Markov's inequality in the proof, simplifying the argument considerably. From the lower bound result in Hoffmann et al. (2015), Theorem 2.1, one should expect that the result holds already for some sufficiently large constant M and that the speed at which the posterior mass of $\{\theta : \|\theta - \theta_0\|^2 > M p_n \log(n/p_n)\}$ converges to zero is $\exp(-C_1 p_n \log(n/p_n))$ for some positive constant C_1. It is well-known

that posterior concentration at rate ϵ_n implies existence of a frequentist estimator with the same rate (cf. Ghosal et al. (2000), Theorem 2.5 for a precise statement). Thus, the rate of contraction around the true mean vector θ_0 must be sharp. This also means that credible sets computed from the posterior cannot be so large as to be uninformative, an effect that, as discussed in the introduction, occurs for the Laplace prior connected to the Lasso. If one wishes to use a credible set centered around the posterior mean, then its radius might still be too small to cover the truth. The first step towards guarantees on coverage is a lower bound on the posterior variance. Such a lower bound was obtained for the horseshoe in Van der Pas et al. (2014), and for priors very closely resembling the horseshoe in Ghosh and Chakrabarti (2015). No such results have been obtained so far for priors on σ_i^2 that have a tail of a different order than $(\sigma_i^2)^{-3/2}$. This is a delicate technical issue that we will not pursue further here.

The results also indicates how to build adaptive procedures. We consider adaptivity to the number of nonzero means, without accounting for the possibly unknown variance of the ε_i, for which a prior of the type suggested for the horseshoe in Carvalho et al. (2010) or an empirical Bayes procedure may be used. The method for adapting to the sparsity does not require explicit knowledge of p_n but in order to get minimax concentration rates, we need to find priors that satisfy the conditions of Theorem 2.1. Consider for example the prior defined as

$$\pi(u) := \frac{1}{u^{3/2}} \frac{\sqrt{\log n}}{n}, \quad \text{for all } u \geq \frac{\sqrt{\log n}}{n}$$

and the remaining mass is distributed arbitrarily on the interval $[0, \sqrt{\log n}/n)$. Thus Condition A holds for any $1 \leq p_n = o(n)$ and thus also Condition 2 and Condition 3. Whenever we impose an upper bound $p_n \leq n^{1-\delta}$ with $\delta > 0$, then also Condition 1 holds and thus Theorem 2.1 follows. This shows that in principle priors can be constructed that adapt over nearly the whole range of possible sparsity levels and lead to some theoretical guarantee. The trick is that a prior that works for an extremely sparse model with $p_n = 1$ also adapts to less sparse models. This requires, however, a lot of prior mass near zero. Such a prior shrinks small non-zero components more than if we first get a rough estimate of the relative sparsity p_n/n and then use a prior that lies on the "boundary" of the conditions in the sense that the both sides in the inequality of Condition 3 are of the same order. An empirical Bayes procedure that first estimates the sparsity was found to work well in Van der Pas et al. (2014), arguing along the lines of Johnstone and Silverman (2004). The sparsity level estimator counts the number of observations that are larger than the 'universal threshold' of $\sqrt{2 \log n}$. Similar results are likely to hold in our setting, as long as the posterior mean is monotone in the parameter that is taken to depend on p_n.

2.2.1 Necessary conditions

The imposed conditions are nearly sharp. To see this, consider the Laplace prior, where each θ_i is drawn independently from a Laplace distribution with parameter λ. It is well-known that the Laplace distribution with parameter λ can be represented as a scale mixture of normals where the mixing density is exponential with parameter λ^2 (cf. Andrews and Mallows (1974) or Park and Casella (2008), Equation (4)). Thus, the Laplace prior fits our

framework (2.1) with $\pi(u) = \lambda^2 e^{-\lambda^2 u}$, for $u \geq 0$. As mentioned in the introduction, the MAP-estimator of this prior is the Lasso but the full posterior does not shrink at the minimax rate. Indeed, Theorem 7 in Castillo et al. (2015) shows that if the true vector is zero, then, the posterior concentration rate has the lower bound n/λ^2 for the squared ℓ^2-norm provided that $1 \leq \lambda = o(\sqrt{n})$. This should be compared to the optimal minimax rate $\log n$ (the rate for sparsity zero is the same as the rate for sparsity $p_n = 1$). Thus, the lower bound shows that the rate is sub-optimal as long as

$$\lambda \ll \sqrt{\frac{n}{\log n}}. \tag{2.7}$$

If $\lambda \gtrsim \sqrt{n/\log n}$, the lower bound is not sub-optimal anymore, but in this case, the non-zero components cannot be recovered with the optimal rate. The lower bound shows that the posterior does not shrink enough if λ is not taken to be huge and thus either Condition 2 or Condition 3 must be violated, as these are the two conditions that guarantee shrinkage of the zero mean coefficients.

Obviously, $\int_0^1 \pi(u)du \geq \int_0^1 e^{-u}du > 0$ for $1 \leq \lambda$ and thus Condition 2 holds. For Condition 3 notice that the integral can be split into the integral $\int_0^1 u\pi(u)du$ plus an integral over $[1,\infty)$ Now, if λ tends to infinity faster than a polynomial order in n then the integral over $[1,\infty)$ is exponentially small in n. Thus Condition 3 must fail because the integral over $\int_{s_n}^1 u\pi(u)du$ is of a larger order than $s_n = n^{-1}\log n$. To see this, observe that for $\lambda \leq \sqrt{n/\log n}$,

$$\int_{s_n}^1 u\lambda^2 e^{-\lambda^2 u}du = \frac{1}{\lambda^2}\int_{s_n\lambda^2}^{\lambda^2} ve^{-v}dv \geq \frac{1}{\lambda^2}\int_1^{\lambda^2} e^{-v}dv \gtrsim \frac{1}{\lambda^2}.$$

Now, we see that Condition 3 fails if and only if (2.7) holds. Indeed, if $\lambda \ll \sqrt{n/\log n}$, then the r.h.s. is of larger order than s_n and if $\lambda \asymp \sqrt{n/\log n}$, then, Condition 3 holds. This shows that this bound is sharp.

In order to state this as a formal result, let us introduce the following modification of Condition 3. Let κ_n denote an arbitrary positive sequence.

Condition 3(κ_n). Let $b_n = \sqrt{\log(n/p_n)}$ and assume that there is a constant C, such that

$$\kappa_n \int_{s_n}^1 u\pi(u)du + \int_1^\infty \Big(u \wedge \frac{b_n^3}{\sqrt{u}}\Big)\pi(u)du + b_n \int_1^{b_n^2} \frac{\pi(u)}{\sqrt{u}}du \leq Cs_n.$$

In particular, we recover Condition 3 for $\kappa_n = 1$.

Theorem 2.2. *Work under model $X^n \sim \mathcal{N}(\theta_0, I_n)$ and assume that the prior is of the form (2.1). For any positive sequence $(\kappa_n)_n$ tending to zero, there exists a prior π satisfying Condition 2 and Condition 3(κ_n) for $p_n = 1$ and a positive sequence $(M_n)_n$ tending to infinity, such that*

$$\mathbb{E}_{\theta_0=0}\Pi\Big(\theta : \|\theta\|_2^2 \leq M_n \log(n) \,\Big|\, X^n\Big) \to 0, \quad \text{as } n \to \infty. \tag{2.8}$$

This theorem shows that the posterior puts asymptotically all mass outside an ℓ^2-ball with radius $M_n \log(n) \gg \log(n)$ and is thus suboptimal. The proof can be found in the appendix.

2.3 Examples

In this section, Conditions 1-3 are verified for the horseshoe-type priors considered by Ghosh and Chakrabarti (2015) (which includes the horseshoe and the normal-exponential gamma), the horseshoe+, the inverse-Gaussian prior, the normal-gamma prior, and the spike-and-slab Lasso. There are, to the best of our knowledge, no existing results yet showing that the horseshoe+, the inverse-Gaussian and the normal-gamma priors lead to posterior contraction at the minimax estimation rate. Posterior concentration for the horseshoe and horseshoe-type priors were already established in Van der Pas et al. (2014) and Ghosh and Chakrabarti (2015), and for the spike-and-slab Lasso in Ročková (2015) . Here, we obtain the same results but thanks to Theorem 2.1 the proofs become extremely short. In addition, we can show that a restriction on the class of priors considered by Ghosh and Chakrabarti (2015) can be removed.

2.3.1 Global-local scale mixtures of normals

In Ghosh and Chakrabarti (2015), the priors under consideration are normal priors with random variances of the form

$$\theta_i \mid \sigma_i^2, \tau^2 \sim \mathcal{N}(0, \sigma_i^2 \tau^2), \quad \sigma_i^2 \sim \pi'(\sigma_i^2), \quad i = 1, \ldots, n,$$

for priors π' with density given by

$$\pi'(\sigma_i^2) = K \frac{1}{(\sigma_i^2)^{a+1}} L(\sigma_i^2), \tag{2.9}$$

where $K > 0$ is a constant and $L : (0, \infty) \to (0, \infty)$ is a non-constant, *slowly varying* function, meaning that there exist $c_0, M \in (0, \infty)$ such that $L(t) > c_0$ for all $t \geq t_0$ and $\sup_{t \in (0,\infty)} L(t) \leq M$. Ghosh and Chakrabarti (2015) prove an equivalent of Theorem 2.1 for these priors, for $a \in [1/2, 1)$ and $\tau = (p_n/n)^\alpha$ with $\alpha \geq 1$.

The horseshoe prior, with $\pi(u) = (\pi\tau)^{-1} u^{-1/2} (1 + u/\tau^2)^{-1}$, is contained in this class of priors, by taking $a = 1/2$, $L(t) = t/(1 + t)$, and $K = 1/\pi$. This class also contains the normal-exponential-gamma priors of Griffin and Brown (2005), for which $\pi(u) = \lambda/\gamma^2 (1 + u/\gamma^2)^{-(\lambda+1)}$ with parameters $\lambda, \gamma > 0$. This class of priors is of the form (2.9) for the choice $\tau = \gamma$, $a = \lambda$ and $L(t) = (t/(1+t))^{1+\lambda}$. In Ghosh and Chakrabarti (2015), it is stated that the three parameter beta normal mixtures, the generalized double Pareto, the inverse gamma and half-t priors are of the form (2.9) as well.

The global-local scale prior is of the form (2.1) with

$$\pi(u) = \frac{K \tau^{2a}}{u^{1+a}} L\Big(\frac{u}{\tau^2}\Big).$$

We assume that the polynomial decay in u is at least of order $3/2$, that is $a \geq \frac{1}{2}$. In particular, the horseshoe lies directly at the boundary in this sense. Depending on a, we allow for different values of τ. If $\frac{1}{2} \leq a < 1$, we assume $\tau^{2a} \leq (p_n/n)\sqrt{\log(n/p_n)}$; if $a = 1$, we assume $\tau^2 \leq p_n/n$; and if $a > 1$, we assume $\tau^2 \leq (p_n/n)\log(n/p_n)$.

Below, we check Conditions 1-3.

Condition 1': It is enough to show that π' is a uniformly regular varying function. Notice that L is uniformly regular varying and satisfies (2.3) with $R = M/c_0$ and $z_0 = t_0$. If two functions are uniformly regular varying, then also their product, and thus π' is uniformly regular varying.

Condition 2: Because of $p_n = o(n)$, $\tau^2 \to 0$. Observe that $u \geq t_0\tau^2$ implies $L(u/\tau^2) \geq c_0$ and thus

$$\int_0^1 \pi(u)du \geq \int_{t_0\tau^2}^{(t_0+1)\tau^2} \pi(u)du \geq \int_{t_0\tau^2}^{(t_0+1)\tau^2} \frac{c_0 K\tau^{2a}}{u^{1+a}}du = \frac{c_0 K}{(t_0+1)^{1+a}}.$$

Condition 3: Since L is bounded in sup-norm by M, and $s_n \geq \tau^2$, we find that $\pi(u) \leq KM\tau^{2a}u^{-1-a}$, for all $u \geq s_n$. With this bound, it is straightforward to verify Condition 3.

Thus, we can apply Theorem 2.1. □

In particular, the posterior concentration theorem holds even more generally than shown by Ghosh and Chakrabarti (2015), as the restriction $a < 1$ can be removed. Thus, for example, we recover Theorem 1.3 of Chapter 1 and in addition, find that the normal-exponential-gamma prior of Griffin and Brown (2005) contracts at at most the minimax rate for $\gamma = p_n/n$ and any $\lambda \geq 1/2$.

2.3.2 The inverse-Gaussian prior

Caron and Doucet Caron and Doucet (2008) propose to use the inverse-Gaussian distribution as prior for σ^2. For positive constants b and τ the variance σ^2 is drawn from an inverse Gaussian distribution with mean $\sqrt{2}\tau$ and shape parameter $\sqrt{2b}$. Thus the prior on the components is of the form (2.1) with

$$\pi(u) = \frac{C_{b,\tau}\tau}{u^{3/2}} e^{-\frac{\tau^2}{u} - bu},$$

where $C_{b,\tau} = e^{2\sqrt{b}\tau}/\sqrt{\pi}$ is the normalization factor. (In the notation of Caron and Doucet (2008), this corresponds to reparametrizing $\gamma = \sqrt{2b}$, $\alpha/n = \sqrt{2}\tau$, and $K = n$ is the dimension of the unknown mean vector.) As τ becomes small the distribution is concentrated near zero. Caron and Doucet (2008) suggests to take τ proportional to $1/n$, and we find that optimal rates can be achieved if $(p_n/n)^K \lesssim \tau \leq (p_n/n)\sqrt{\log(n/p_n)}$ for some $K > 1$.

Below we verify Condition 1 and Condition A, which together imply Theorem 2.1. The inverse-Gaussian prior does not fit within the class considered by Ghosh and Chakrabarti (2015), because of the additional exponential factors.

Condition 1: For $u \geq 1$, $e^{-1} \leq e^{-\tau^2/u} \leq 1$. Thus, $u \mapsto e^{-\tau^2/u}$ is uniformly regular varying with constants $R = e$ and $z_0 = 1$. Since products of uniformly regular varying functions are again uniformly regular varying, we can write $\pi(u) = L_n(u)e^{-bu}$ with L_n uniformly regular varying.

For $u \geq 1$, $\pi(u) \geq \pi^{-1/2} e^{-1} \tau u^{-3/2} e^{-bu}$, using the explicit expression for the constant $C_{b,\tau}$. Thus, (2.4) holds with $b' > b, K = \alpha, z_* = 1$, and C' a sufficiently large constant.

Condition A: Observe that $\pi(u) \leq C_{b,1} \tau u^{-3/2}$.

Hence, the statement of Theorem 2.1 follows. □

2.3.3 The horseshoe+ prior

The horseshoe+ prior was introduced by Bhadra et al. (2015). It is an extension of the horseshoe including an additional latent variable. A Cauchy random variable with parameter λ that is conditioned to be positive is said to be half-Cauchy and we write $C^+(0,\lambda)$ for its distribution. The horseshoe+ prior can be defined via the hierarchical construction

$$\theta_i \mid \sigma_i \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i \mid \eta_i, \tau \sim C^+(0, \tau\eta_i), \quad \eta_i \sim C^+(0,1).$$

and should be compared to the horseshoe prior

$$\theta_i \mid \sigma_i \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i \mid \tau \sim C^+(0, \tau).$$

The additional variable η_i allows for another level of shrinkage, a role which falls solely to τ in the horseshoe prior. In Bhadra et al. (2015), the claim is made that the horseshoe+ is an improvement over the horseshoe in several senses, but no posterior concentration results are known so far. With Theorem 2.1, we can show that the horseshoe+ enjoys the same upper bound on the posterior contraction rate as the horseshoe, if $(p_n/n)^K \lesssim \tau \lesssim (p_n/n)(\log(n/p_n))^{-1/2}$, for some $K > 1$.

The horseshoe+ prior is of the form (2.1) with

$$\pi(u) = \frac{\tau}{\pi^2} \frac{\log(u/\tau^2)}{(u-\tau^2)u^{1/2}}.$$

Below, we verify Conditions 1-3.

Condition 1: Write $\pi(u) = L_n(u)$, that is, $b = 0$. Let us show that L_n is uniformly regular varying. For that define $u_0 := 2$. For $u > u_0$, and $\tau^2 \leq 1$ we have $u/2 \leq u - \tau^2 \leq u$, thus

$$\frac{1}{2} a^{-3/2} \frac{\log(u/\tau^2) + \log(a)}{\log(u/\tau^2)} \leq \frac{\pi(au)}{\pi(u)} \leq 2a^{-3/2} \frac{\log(u/\tau^2) + \log(a)}{\log(u/\tau^2)}.$$

Since

$$1 \leq \frac{\log(u/\tau^2) + \log(a)}{\log(u/\tau^2)} \leq 2,$$

L_n is regular varying. To check the second part of the assumption, observe that $\pi(u) \geq \pi^{-1} \tau u^{-3/2} \log(u/\tau^2)$. For any $K > \alpha$ and any $b' > 0$,

$$\pi(u) e^{b'u} \gtrsim \tau \log(1/\tau) \geq \left(\frac{p_n}{n}\right)^K, \quad \text{for all } u \geq u_0.$$

Thus, Condition 1 holds.

Condition 2: Observe that

$$\int_0^1 \pi(u) du \geq \frac{\tau}{\pi^2} \int_0^{\tau^2/2} \frac{\log(\tau^2/u)}{(\tau^2-u)u^{1/2}} du \geq \frac{\tau}{\pi^2} \frac{1}{(\tau^2/2)^{3/2}} \cdot \frac{\tau^2}{2} \log \tfrac{1}{2} \gtrsim 1.$$

Condition 3: For any $u \geq s_n$ we can use $(u - \tau^2) \geq u/2$. This shows that

$$\pi(u) \leq \frac{\tau \log(u)}{u^{3/2}} + \frac{\tau \log(1/\tau^2)}{u^{3/2}}, \quad \text{for all } u \geq s_n.$$

In particular, $\pi(u) \lesssim \tau \log(n/p_n)/u^{3/2}$ for $s_n \leq u \leq b_n^2$. For the integral on $[b_n^2, \infty)$, we use that $\frac{d}{du} - (\log(u) + 1)/u = \log(u)/u^2$. Together, Condition 3 follows thanks to $\tau \lesssim (p_n/n)/\sqrt{\log(n/p_n)}$.

Thus, Theorem 2.1 can be applied. □

2.3.4 Normal-gamma prior

The normal-gamma prior, discussed by Caron and Doucet (2008) and Griffin and Brown (2010), takes the following form for shape parameter $\tau > 0$ and rate parameter $\beta > 0$:

$$\pi(u) = \frac{\beta^\tau}{\Gamma(\tau)} u^{\tau-1} e^{-\beta u} = \frac{\tau \beta^\tau}{\Gamma(\tau+1)} u^{\tau-1} e^{-\beta u}.$$

In Griffin and Brown (2010), it is observed that decreasing τ leads to a distribution with a lot of mass near zero, while preserving heavy tails. This is also illustrated in the right-most panels of Figure 2.1. The class of normal-gamma priors includes the double exponential prior as a special case, with $\tau = 1$. We now show that the normal-gamma prior satisfies the conditions of Theorem 2.1 for any fixed β, and for any $(p_n/n)^K \lesssim \tau \lesssim (p_n/n)\sqrt{\log(n/p_n)} \leq 1$ for some fixed K.

Below, we check Conditions 1-3.

Condition 1: We define $L_n(u) = \frac{\beta^\tau}{\Gamma(\tau)} u^{\tau-1}$, so $\pi(u) = L_n(u) e^{-bu}$ with $b = \beta$. Note that since $\tau \to 0$, we have that there exist a constant C such that $C^{-1} \leq \beta^\tau \leq C$. We now prove that L_n is regular varying. We have

$$\frac{L_n(au)}{L_n(u)} = a^{\tau-1}.$$

and thus for all $a \in [1,2]$, $a^{-1} \leq L_n(au)/L_n(u) \leq 1$. In addition for $u > u_* := 1$ we have, using $\Gamma(\tau+1) \geq \Gamma(1) = 1$,

$$L_n(u) = \frac{\tau \beta^\tau}{\Gamma(\tau+1)} u^{\tau-1} \geq \frac{(\beta \wedge 1)\tau}{\Gamma(2) u} \gtrsim \left(\frac{p_n}{n}\right)^K \frac{1}{u},$$

implying $\pi(u) = L_n(u) u^{-1} e^{-\beta u} \gtrsim (p_n/n)^K e^{-2\beta u}$. Thus Condition 1 is satisfied.

Condition 2:

$$\int_0^1 \pi(u) du \geq \frac{(\beta \wedge 1) e^{-bu} \tau}{\Gamma(2)} \int_0^1 u^{\tau-1} du = \frac{(\beta \wedge 1) e^{-bu}}{\Gamma(2)} \gtrsim 1.$$

Condition 3: Notice that $\pi(u) \leq (\beta \vee 1)\tau u^{\tau-1}$, for all $u \leq 1$. For $u \geq 1$, we find $\pi(u) \leq (\beta \vee 1)\tau e^{-\beta u}$. Since $e^{-\beta u}$ decays faster than any polynomial power of u, we see that Condition 3 holds thanks to $b_n \tau \lesssim s_n$.

Thus, we can apply Theorem 2.1.

In Griffin and Brown (2010), it is discussed that the extra modelling flexibility afforded by generalizing the double exponential prior to include the parameter τ is essential, and indeed the double exponential ($\tau = 1$) does not allow a dependence on p_n and n such that our conditions are met.

2.3.5 Spike-and-slab Lasso prior

The spike-and-slab Lasso prior was introduced by Ročková (2015). It may be viewed as a continuous version of the usual spike-and-slab prior with a Laplace slab, as studied in Castillo et al. (2015); Castillo and Van der Vaart (2012), where the spike component has been replaced by a very concentrated Laplace distribution. Recent theoretical results, including posterior concentration at the minimax rate, have been obtained in Ročková (2015). Here, we recover Corollary 6.1 of Ročková (2015).

For a fixed constant $a > 0$ and a sequence $\tau \to 0$, we define the spike-and-slab Lasso as prior of the form (2.1) with hyperprior

$$\pi(u) = \omega a e^{-au} + (1-\omega)\frac{1}{\tau}e^{-\frac{u}{\tau}}, \quad u > 0 \tag{2.10}$$

on the variance. Recall that the Laplace distribution with parameter λ is a scale mixture of normals where the mixing density is exponential with parameter λ^2. Applied to model (2.1), the prior on θ_i is thus a mixture of two Laplace distributions with parameter $\sqrt{a}$ and $\tau^{-1/2}$ and mixing weights ω and $1-\omega$, respectively and this justifies the name.

We now prove that the prior satisfies the conditions of Theorem 2.1 for mixing weights satisfying $(p_n/n)^K \leq \omega \leq (p_n/n)\sqrt{\log(n/p_n)} \leq \frac{1}{2}$, for some $K > 1$ and $\tau = (p_n/n)^\alpha$ with $\alpha \geq 1$.

Condition 1: To prove that Condition 1 holds we rewrite the prior π as

$$\pi(u) = e^{-au}\left(a\omega + \frac{1-\omega}{\tau}e^{-u(\frac{1}{\tau}-a)}\right) =: e^{-au}L_n(u)$$

For n large enough, we have $1/\tau - a > 1/(2\tau)$. For all $u > 1$ and for $C > 0$ a constant depending only on K and α,

$$\frac{1-\omega}{\tau}e^{-u(\frac{1}{\tau}-a)} \leq \frac{1}{\tau}e^{-\frac{1}{2\tau}} \leq C\tau^{\frac{K}{\alpha}} \leq C\omega.$$

Hence, for sufficiently large n, $a\omega \leq L_n(u) \leq (a+C)\omega$ for all $u \geq 1$. Thus L_n is regular varying with $u_0 = 1$. Since also $\pi(u) \geq a\omega e^{-au}$ and $\omega \geq (p_n/n)^K$, Condition 1 holds.

Condition 2: $\int_0^1 \pi(u)du \geq (1-\omega)\int_0^\tau \frac{1}{\tau}e^{-\frac{u}{\tau}}du = (1-\omega)(1-e^{-1})$.

Condition 3: We might split the two mixing components in (2.10) and write $\pi =: \pi_1 + \pi_2$. To verify the condition for the first component π_1, we use that $e^{-au} \leq 1$ for $u \leq 1$ and that e^{-au} decays faster than any polynomial for $u > 1$. In order that Condition 3 is satisfied, we need thus $\omega \lesssim (p_n/n)\sqrt{\log(n/p_n)}$. For π_2, there exists a constant C such that $\pi_2(u) \leq C\tau/u^2$ for all $u \geq s_n$, due to $s_n \geq \tau$. Straightforward computations show that π_2 satisfies Condition 3 since $\tau \leq p_n/n$.

Thus, we can apply Theorem 2.1. □

2.4 Simulation study

To illustrate the point that our conditions are very sharp, we compute the average square loss for four priors that do not meet our conditions, and compare them with two of the examples from Section 2.3.

The two priors considered in this simulation study that do meet the conditions are the horseshoe and the normal-gamma priors, both with $\tau = p_n/n$. The four priors that do not meet the conditions are the Lasso (Laplace prior) with $\lambda = 1$ and $\lambda = 2n/\log n$ (see Section 2.3.4), and two priors of the form (2.9) of Section 2.3.1 with $a = 0.1$ and $a = 0.4$, $L(u) = e^{-1/u}$ and density,

$$\pi(u) \propto u^{-(1+a)} e^{-\tau^2/u},$$

and we take $\tau = p_n/n$. This prior will be referred to as a $GC(a)$ prior hereafter. Note that π does not meet our conditions, as explained in Section 2.3.1.

For each of these priors, we sample from the posterior distribution using a Gibbs Sampling algorithm, following the one proposed for the horseshoe prior by Carvalho et al. (2010). To do so, we first compute the full conditional distributions

$$p(\beta|X,\sigma^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-\frac{1}{2\hat{\sigma}^2}(\beta-\hat{\beta})^2}$$
$$p(\sigma^2|X,\beta) \propto (\sigma^2)^{-1/2} e^{-\frac{\beta^2}{2\sigma^2}} \pi(\sigma^2),$$

where $\hat{\sigma}^2 = \sigma^2/(1+\sigma^2)$ and $\hat{\beta} = X\sigma^2/(1+\sigma^2)$. The only difficulty is thus sampling from $p(\sigma^2|X,\beta)$. For the horseshoe prior we follow the approach proposed by Carvalho et al. (2010). We apply a similar method for the normal-gamma prior using the approach proposed by Damien et al. (1999). Sampling from the $GC(a)$ priors is even simpler given that in this case $p(\sigma|X,\beta)$ is an inverse gamma. We compute the mean integrated squared error (MISE) on 500 replicates of simulated data of size $n = 100, 250, 500, 1000$. The MISE is equal to $\mathbb{E}_{\theta_0} \sum_i [(\widehat{\theta}_i - \theta_{0i})^2 + \mathrm{var}(\theta_i \mid X)]$. For each n, we fix the number of nonzero means at $p_n = 10$, and take the nonzero coefficients equal to $5\sqrt{2\log n}$. This value is well past the 'universal threshold' of $\sqrt{2\log n}$, and thus the signals should be relatively easy to detect. For each data set, we compute the posterior square loss using 5000 draws from the posterior with a burn-in of 20%.

The results are presented in Figure 2.2, for all means together and separately for the nonzero and zero means. Given that $p_n = 10$ is fixed, if the posterior contracts at the minimax rate, then the integrated square loss should be linear in $\log n$. However, we see that for both Laplace priors and the $GC(a = 0.1)$ priors, and less so for the $GC(a = 0.4)$ prior, the slope of the loss grows with n, when it remains steady for the other two considered priors. In addition, we see the expected trade-off for the two choices of the tuning parameter λ for the Lasso. A large value of λ results in strong shrinkage and thus low MISE on the zero means, but very high MISE on the nonzero means, while a small value of λ leads to barely any shrinkage, and we observe a relatively low MISE on the nonzero means but a high MISE on the zero means. The $GC(a)$ prior with $a = 0.1$ does not perform well, because it undershrinks. The same effect is visible for $a = 0.4$, but less so. The

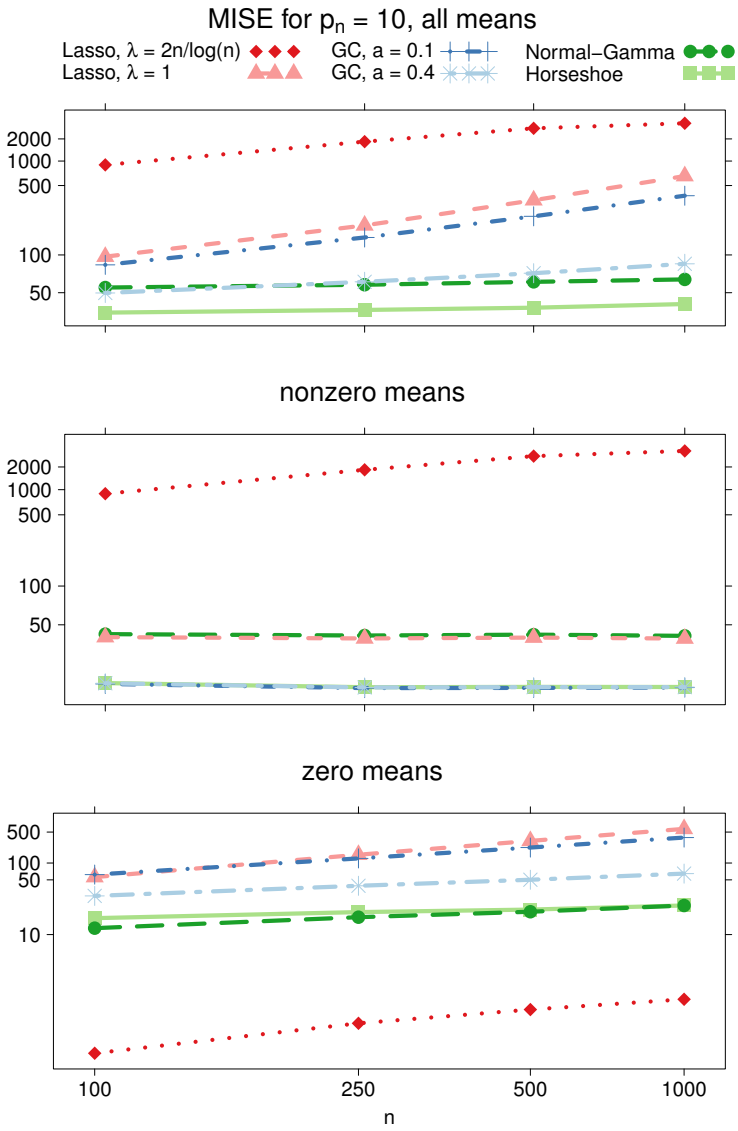

Figure 2.2: The logarithm of the integrated square loss for the Lasso (Laplace) with $\lambda = 2n/\log n$ and $\lambda = 1$, the GC priors of Ghosh and Chakrabarti (2015) discussed in section 2.3.1 with $a = 0.1$ and $a = 0.4$, the normal-namma and horseshoe priors plotted against $\log\log n$, computed on 500 replicates of the data for each value of n. From top to bottom: MISE for all means, for only the $p_n = 10$ nonzero means, and for the $(n - p_n)$ zero means. The axis labels refer to the original, non-log-transformed scale.

normal-gamma and horseshoe priors both have low MISE on the zero and nonzero means; the horseshoe outperforms the normal-gamma because it shrinks the nonzero means less.

These results suggest that the horseshoe and normal-gamma strike a better balance between shrinking the zero means without affecting the nonzero means than the four priors that do not meet our conditions, leading to lower risk and illustrating that our conditions are very sharp.

2.5 Discussion

Our main theorem, Theorem 2.1, expands the class of shrinkage priors with theoretical guarantees for the posterior contraction rate. Not only can it be used to obtain the optimal posterior contraction rate for the horseshoe+, the inverse-Gaussian and normal-gamma priors, but the conditions provide some characterization of properties of sparsity priors that lead to desirable behaviour. Essentially, the tails of the prior on the local variance should be at least as heavy as Laplace, but not too heavy, and there needs to be a sizable amount of mass around zero compared to the amount of mass in the tails, in particular when the underlying mean vector grows to be more sparse.

In Polson and Scott (2010) global-local scale mixtures of normals like (2.5) are discussed, with a prior on the parameter τ^2. Their guidelines are twofold: the prior on the local variance σ_i^2 should have heavy tails, while the prior on the global variance τ^2 should have substantial mass around zero. They argue that any prior on σ_i^2 with an exponential tail will force a tradeoff between shrinking the noise towards zero and leaving the large nonzero means unshrunk, while the shrinkage of large signals will go to zero when a prior with a polynomial tail is chosen. This matches the intuition behind our conditions, with the remark that exponential tails *are* possible, but they should not be lighter than Laplace.

Besides the three discussed goals of recovery, uncertainty quantification, and computational simplicity, we might have mentioned a fourth: performing *model selection* or *multiple testing*. Priors of the type studied in this paper are not directly applicable for this goal, as the posterior mean will, with probability one, not be exactly equal to zero. A model selection procedure can be constructed however, for example by thresholding using the observed values of m_{x_i}: if m_{x_i} is larger than some constant, we consider the underlying parameter to be a signal, and otherwise we declare it noise. Such a procedure was proposed for the horseshoe by Carvalho et al. (2010), and was shown to enjoy good theoretical properties by Datta and Ghosh (2013). Similar results were found for the horseshoe+ (Bhadra et al., 2015). The same thresholding procedure, and similar analysis methods, may prove to be fruitful for the more general prior (2.1).

2.6 Proofs

This section contains the proofs of Theorem 2.1 and Theorem 2.2, followed by the statement and proofs of the supporting Lemmas. The proof of Theorem 2.1 follows the same structure as that of Theorem 1.3 in Chapter 1, but requires more general methods to bound the integrals involved in the proof.

In the course of the proofs, we use the following two transformations of π,

$$g(z) = \frac{1}{z^2}\pi\left(\frac{1-z}{z}\right) \quad \text{and} \quad h(z) = \frac{1}{(1-z)^{3/2}}\pi\left(\frac{z}{1-z}\right). \tag{2.11}$$

The function g is a density on $[0,1]$, resulting from transforming the density π on σ_i^2 to a density for $z = (1+\sigma_i^2)^{-1}$. The function h is a rescaled version of π.

Lemma 2.3. *Condition 1' implies Condition 1.*

Proof. Observe that $\pi(u) = \widetilde{\pi}(u/\tau^2)/\tau^2$. Since by assumption $\widetilde{\pi}$ is uniformly regular varying, (2.3) holds for some constants R and u_0 which do not depend on n. To check the first part of Condition 1, it is enough to see that $\widetilde{\pi}(\cdot/\tau^2)$ is uniformly regular varying as well and satisfies (2.3) with the same constants as $\widetilde{\pi}$.

It remains to prove a lower bound (2.4). Thanks to $\tau^2 \leq 1$ and Lemma 2.5, for any $u \geq u_* := u_0$, $\widetilde{\pi}(u/\tau^2) \geq \widetilde{\pi}(u_0)(\tau^2 u_0/2u)^{\log_2 R}$. This implies the lower bound (2.4) with $K = 2\alpha \log_2 R$, $b' > 0$, and C' a sufficiently large constant. □

Proof of Theorem 2.1. Applying Lemma 2.7 gives under Condition 1, $\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i}(\theta_i - \widehat{\theta_i})^2 \lesssim p_n \log(n/p_n)$ and $\sum_{i:\theta_i\neq 0} \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i \mid X_i) \lesssim p_n \log(n/p_n)$. These inequalities combined with Markov's inequality prove the first two statements of the theorem. Similarly, under Condition 2 and Condition 3, we obtain from Lemma 2.8 and Lemma 2.9, $\mathbb{E}_\theta \sum_{i:\theta_i=0} \widehat{\theta_i^2} \leq n\mathbb{E}_0(Xm_X)^2 \lesssim p_n \log(n/p_n)$ and $\sum_{i:\theta_i=0} \mathbb{E}_0 \operatorname{var}(\theta_i \mid X_i) \lesssim p_n \log(n/p_n)$. Together with Markov's inequality, this proves the third and fourth statement of the theorem. □

Proof of Theorem 2.2. Without loss of generality, we can take κ_n such that $\kappa_n \geq n^{-1/4}$ for all n. Consider the prior, where θ_i is drawn from the Laplace density with parameter $\lambda = \sqrt{s_n/\kappa_n}$. This prior is of the form (2.1) with $\pi(u) = \lambda^2 e^{-\lambda^2 u}$ (cf. Section 2.2.1). Theorem 7 in Castillo et al. (2015) shows that (2.8) holds with $M_n = 1/\kappa_n \to \infty$. Thus it remains to prove that π satisfies Condition 2 and Condition 3(κ_n).

Condition 2 follows immediately. For Condition 3(κ_n) observe that due to $\kappa_n \geq n^{-1/4}$, $\lambda \geq n^{1/4}/\sqrt{\log n}$. Splitting the integral $\int_0^{\lambda^2} = \int_0^1 + \int_1^{\lambda^2}$, we find $\kappa_n \int_{s_n}^1 u\pi(u)du \leq \kappa_n \int_0^1 u\lambda^2 e^{-\lambda^2 u}du \leq \kappa_n\lambda^{-2}\int_0^{\lambda^2} ve^{-v}dv \lesssim \kappa_n\lambda^{-2} = s_n$. Also, $\int_1^{b_n^2} u\pi(u)du = \lambda^{-2}\int_{\lambda^2}^{b_n^2\lambda^2} ve^{-v}dv \leq b_n^2 e^{-\lambda^2} = o(s_n)$ and $b_n^3\int_1^\infty \pi(u)/\sqrt{u}du \leq b_n^3\int_1^\infty \pi(u)du \leq b_n^3 e^{-\lambda^2} = o(s_n)$. Hence, Condition 3($\kappa_n$) holds and this completes the proof. □

Lemma 2.4. *The posterior variance can be written as*

$$\operatorname{var}(\theta \mid x) = m_x - (xm_x - x)^2 + x^2\frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z}dz}{\int_0^1 h(z)e^{\frac{x^2}{2}z}dz} \tag{2.12}$$

and bounded by

$$\operatorname{var}(\theta \mid x) \leq 1 + x^2\frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z}dz}{\int_0^1 h(z)e^{\frac{x^2}{2}z}dz} \quad \textit{and} \quad \operatorname{var}(\theta \mid x) \leq m_x + x^2 m_x. \tag{2.13}$$

Proof. By Tweedie's formula (Robbins, 1956), the posterior variance for θ_i given an observation x_i is equal to $1 + (d^2/dx^2) \log p(x)|_{x=x_i}$, where $p(x_i)$ is the marginal distribution of x_i. Computing

$$p(x) = \int_0^1 \frac{1}{\sqrt{2\pi}} (1-z)^{-3/2} e^{-\frac{x^2}{2}(1-z)} \pi\left(\frac{z}{1-z}\right) dz,$$

taking derivatives with respect to x, and substituting $h(z) = (1-z)^{-3/2}\pi(z/(1-z))$ gives

$$\begin{aligned}\operatorname{var}(\theta \mid x) = 1 + x^2 \frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} - \frac{\int_0^1 (1-z) h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \\ - x^2 \left(\frac{\int_0^1 (1-z) h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \right)^2.\end{aligned}$$

From that we can derive (2.12) noting that the third term on the r.h.s. is $1 - m_x$. The last display also implies the first inequality in (2.13). Representation (2.12) together with the trivial bound $(1-z)^2 \leq (1-z)$ for $z \in [0,1]$ yields

$$x^2 \frac{\int_0^1 (1-z)^2 h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} \leq x^2 \frac{\int_0^1 (1-z) h(z) e^{\frac{x^2}{2}z} dz}{\int_0^1 h(z) e^{\frac{x^2}{2}z} dz} = x^2(1-m_x).$$

Combined with (2.12), we find $\operatorname{var}(\theta \mid x) \leq m_x - x^2 m_x^2 + x^2 m_x \leq m_x + x^2 m_x$. □

Lemma 2.5. *Suppose that L is uniformly regular varying. If R and u_0 are chosen such that* (2.3) *holds, then, for any $a \geq 1$, and any $u \geq u_0$,*

$$L(u) \leq (2a)^{\log_2 R} L(au),$$

where $\log_2$ *denotes the binary logarithm.*

Proof. Write $a = 2^r b$ with r a non-negative integer and $1 \leq b < 2$. By assumption (2.3) holds for some R and u_0. We apply the upper bound (2.3) repeatedly and obtain for $a \geq 1$, $L(u) \leq RL(2u) \leq \ldots \leq R^r L(2^r u) \leq R^{r+1} L(au)$. Since $R^{r+1} = (2^{r+1})^{\log_2 R} \leq (2a)^{\log_2 R}$, the result follows. □

Lemma 2.6. *Assume that L is uniformly regular varying and satisfies* (2.3) *with R and u_0. Then, the shifted function $L(\cdot - 1)$ is also uniformly regular varying with constants R^3 and $u_0 \vee 2$.*

Proof. Write

$$\frac{L(az-1)}{L(z-1)} = \frac{L(az-1)}{L(az)} \cdot \frac{L(az)}{L(z)} \cdot \frac{L(z)}{L(z-1)}.$$

For $z \geq z_0 \vee 2$ we apply (2.3) to each of the three fractions and this completes the proof. □

The following lemma states that if the density g can be decomposed as a product of a function that is uniformly varying and possibly n dependent, and a factor of the form $z \mapsto e^{-bz}$, then the posterior recovers the size of the non-zero components of θ with the minimax estimation rate, provided that the n dependence is of the right order.

Lemma 2.7. *If Condition 1 holds, there exists a constant C, which is independent of n, such that*

$$\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i}(X_i m_{X_i} - \theta_i)^2 \leq C p_n \log(en/p_n), \tag{2.14}$$

and

$$\sum_{i:\theta_i \neq 0} \mathbb{E}_{\theta_i} \operatorname{var}(\theta_i | X_i) \leq C p_n \log(en/p_n). \tag{2.15}$$

Proof. We prove the two statements separately. The main argument is a careful analysis of the integral representation

$$|x(m_x - 1)| = |x| \frac{\int_0^1 e^{-\frac{x^2}{2} z} z^{-1/2} \pi\Big(\frac{1}{z} - 1\Big) dz}{\int_0^1 e^{-\frac{x^2}{2} z} z^{-3/2} \pi\Big(\frac{1}{z} - 1\Big) dz} = |x| \frac{\int_0^1 e^{-\frac{x^2}{2} u} u^{3/2} g(u) du}{\int_0^1 e^{-\frac{x^2}{2} u} u^{1/2} g(u) du}$$

(cf. (2.2) and (2.11)). Throughout the remaining proof, let C_1 be a generic constant which is independent of n and which might change from line to line. Without loss of generality, we may assume that $u_0 \geq 2$ in Condition 1.

Proof of (2.14)*:* It is enough to show $\sup_{x>0} |x(m_x - 1)| \lesssim 1 + \sqrt{\log(n/p_n)}$. It is thus enough to consider the sup over $|x| > T_0 := 2 + 2(u_0 \vee u_*) + \sqrt{8u_0 K \log(n/p_n)}$, since otherwise, we simply use $|x(m_x - 1)| \leq |x|$.

For $0 \leq a < b \leq 1$, write $I(a,b) = \int_a^b e^{-\frac{x^2}{2} u} u^{3/2} g(u) du / \int_0^1 e^{-\frac{x^2}{2} u} u^{1/2} g(u) du$ and for $b \leq a$, set $I(a,b) = 0$. We need to prove that

$$I(0,1) = I\Big(0, \tfrac{2b+4}{|x|}\Big) + I\Big(\tfrac{2b+4}{|x|}, \tfrac{1}{u_0}\Big) + I\Big(\tfrac{1}{u_0}, 1\Big) =: (I) + (II) + (III) \lesssim \frac{1}{|x|}.$$

Bound for (I) : Obviously, $I(0,v) \leq v$ for all $v \in (0,1]$. Thus, $I\Big(0, \frac{2b+4}{|x|}\Big) \leq C_1/|x|$.

Bound for (II) : We first derive a lower bound for the denominator. Recall that by Condition 1, $\pi(u) = L_n(u) e^{-bu}$. Define $\widetilde{L}_n = L_n(\cdot - 1)$ and observe that due to $|x| \geq 2u_0$ we can use Lemma 2.6 and substitute $v = u|x|/2$ to obtain

$$\begin{aligned}
\int_0^1 e^{-\frac{x^2}{2} u} u^{-3/2} \pi\Big(\tfrac{1}{u} - 1\Big) du &\geq \int_{1/|x|}^{2/|x|} e^{-\frac{x^2}{2} u} u^{-3/2} \widetilde{L}_n\Big(\tfrac{1}{u}\Big) e^{-\frac{b}{u} + b} du \\
&\geq \frac{1}{4} e^{b - (1+b)|x|} |x|^{3/2} \int_{1/|x|}^{2/|x|} \widetilde{L}_n\Big(\tfrac{1}{u}\Big) du \\
&= \frac{1}{4} e^{b - (1+b)|x|} |x|^{1/2} 2 \int_{1/2}^{1} \widetilde{L}_n\Big(\tfrac{1}{v} \cdot \tfrac{|x|}{2}\Big) dv
\end{aligned} \tag{2.16}$$

$$\geq \frac{1}{4R^3} e^{b-(1+b)|x|} |x|^{1/2} \widetilde{L}_n\Big(\tfrac{|x|}{2}\Big). \tag{2.17}$$

For the numerator, using Lemma 2.5 with $u = |x|/v$ and $a = v/2$,

$$\begin{aligned}
&\int_{(2b+4)/|x|}^{u_0^{-1}} e^{-\frac{x^2}{2}u} u^{-1/2} \pi\Big(\tfrac{1}{u} - 1\Big) du \\
&= \sum_{k=1}^{\infty} \int_{(2b+4+k-1)/|x|}^{(2b+4+k)/|x|} e^{-\frac{x^2}{2}u} u^{-1/2} \widetilde{L}_n\Big(\tfrac{1}{u}\Big) e^{b-\frac{b}{u}} \mathbf{1}(u \leq u_0^{-1}) du \\
&\leq e^b \sum_{k=1}^{\infty} e^{-\frac{|x|}{2}(2b+4+k-1)} \Big(\frac{|x|}{2b+4+k-1}\Big)^{1/2} \int_{(2b+4+k-1)/|x|}^{(2b+4+k)/|x|} \widetilde{L}_n\Big(\tfrac{1}{u}\Big) \mathbf{1}(u \leq u_0^{-1}) du \\
&\leq e^b \sum_{k=1}^{\infty} e^{-\frac{|x|}{2}(2b+2+k)} |x|^{-1/2} \int_{2b+4+k-1}^{2b+4+k} \widetilde{L}_n\Big(\tfrac{|x|}{v}\Big) \mathbf{1}\Big(v \leq \tfrac{|x|}{u_0}\Big) dv \\
&\leq e^{-|x|(b+1)} |x|^{-1/2} \widetilde{L}_n\Big(\tfrac{|x|}{2}\Big) e^b \sum_{k=1}^{\infty} e^{-\frac{|x|}{2}k} (2b+4+k)^{3\log_2 R}.
\end{aligned}$$

The sum $\sum_{k=1}^{\infty} e^{-\frac{|x|}{2}k}(2b+4+k)^{3\log_2 R}$ is bounded for $|x| > T_0$. Since by assumption, R does not depend on n, we find $I\Big(\frac{2b+4}{|x|}, \frac{1}{u_0}\Big) \leq C_1/|x|$.

Bound for (III) : Since g is a density, we obtain $\int_{u_0^{-1}}^{1} e^{-\frac{x^2}{2}u} u^{3/2} g(u) du \leq e^{-x^2/(2u_0)}$. For the denominator, we find using (2.17), $|x| \geq 2 + 2u_*$, and Condition 1,

$$\begin{aligned}
\int_0^1 e^{-\frac{x^2}{2}u} u^{-3/2} \pi\Big(\tfrac{1}{u} - 1\Big) du &\geq \frac{1}{4R^3} e^{-(1+\frac{b}{2})|x|} |x|^{1/2} \pi\Big(\tfrac{|x|}{2} - 1\Big) \\
&\geq \frac{1}{4R^3 C'} \Big(\frac{p_n}{n}\Big)^K e^{-(1+b+b')|x|} |x|^{1/2}.
\end{aligned}$$

Combining this with the upper bound and $(1+b+b')|x| \leq (1+b+b')^2 u_0 + x^2/(4u_0)$ gives

$$I\Big(\tfrac{1}{u_0}, 1\Big) \leq 4C' R^3 \Big(\tfrac{n}{p_n}\Big)^K |x|^{-1/2} e^{(1+b+b')^2 u_0} e^{-x^2/(4u_0)}.$$

Using that $x \mapsto |x|^{1/2} e^{-x^2/(8u_0)}$ is bounded and $|x| > T_0$ yields $I\Big(\frac{1}{u_0}, 1\Big) \leq C_1/|x|$.

The result for (2.14) follows by combining the bounds $(I) - (III)$.

Proof of (2.15)*:* Recall that (2.13) uses $h(u) = (1-u)^{-3/2}\pi(u/(1-u))$. With (2.11), $h(1-u) = u^{-3/2}\pi((1-u)/u) = u^{1/2} g(u)$. Therefore, we find

$$\operatorname{var}(\theta|x) \leq 1 + x^2 \frac{\int_0^1 e^{-\frac{x^2}{2}u} u^{5/2} g(u) du}{\int_0^1 e^{-\frac{x^2}{2}u} u^{1/2} g(u) du}.$$

Arguing as for (2.14) completes the proof. □

Next, we provide the technical lemmas establishing the rate for the zero coefficients. Recall that $s_n = (p_n/n)\log(n/p_n)$ and define

$$q_n := \frac{p_n}{n} \sqrt{\log(n/p_n)}. \tag{2.18}$$

Suppose that Condition 2 and Condition 3 hold with constants c and C, respectively. With (2.2),

$$\begin{aligned} m_x &:= \frac{\int_0^\infty \frac{u}{(1+u)^{3/2}} e^{\frac{x^2 u}{2+2u}} \pi(u) du}{\int_0^\infty \frac{1}{(1+u)^{1/2}} e^{\frac{x^2 u}{2+2u}} \pi(u) du} \\ &\le s_n + \frac{\sqrt{2}}{c} \int_{s_n}^\infty \frac{u e^{\frac{x^2 u}{2+2u}}}{(1+u)^{3/2}} \pi(u) du \\ &\le s_n \Big(1 + \frac{\sqrt{2}C}{c} e^{\frac{x^2}{4}}\Big) + \frac{\sqrt{2}}{c} \int_1^\infty \frac{u e^{\frac{x^2 u}{2+2u}}}{(1+u)^{3/2}} \pi(u) du \\ &\le s_n \Big(1 + \frac{\sqrt{2}C}{c} e^{\frac{x^2}{4}}\Big) + \frac{\sqrt{8}C}{c} q_n e^{\frac{x^2}{2}}, \end{aligned} \tag{2.19}$$

where for the last inequality, we split the integral $\int_1^\infty = \int_1^{\log(n/p_n)} + \int_{\log(n/p_n)}^\infty$ and used Condition 3 twice. These inequality will be very useful for the proofs below. For the variance bound, the last bound is not sharp enough and we need to work with the upper bound induced by the second inequality.

Lemma 2.8. *Work under Condition 2 and Condition 3. Then,*

$$\mathbb{E}_0 (X m_X)^2 \lesssim \frac{p_n}{n} \log(n/p_n).$$

Proof. Let q_n be as in (2.18) and set $a_n := \sqrt{2 \log(1/q_n)}$. Decompose

$$\mathbb{E}_0 (X m_X)^2 = \mathbb{E}_0 (X m_X)^2 \mathbf{1}\{|X| \le a_n\} + \mathbb{E}_0 (X m_X)^2 \mathbf{1}\{|X| > a_n\} =: I_1 + I_2.$$

To bound the term I_1, (2.19) and $x^2 e^{x^2/2} \le \frac{d}{dx}[x e^{x^2/2}]$ yield

$$I_1 \lesssim s_n^2 \int_{-a_n}^{a_n} x^2 dx + q_n^2 \int_{-a_n}^{a_n} x^2 e^{x^2/2} dx \lesssim s_n^2 a_n^3 + q_n^2 a_n e^{a_n^2/2}.$$

There is a constant only depending on K such that $x^2 \log^K(1/x) \le C_K x$ for all $x \le 1$. Thus, $I_1 \lesssim (p_n/n) \log(n/p_n)$.

In order to bound I_2, we use $m_x \le 1$, $\frac{d}{dx}[-x e^{-x^2/2}] = -e^{-x^2/2} + x^2 e^{-x^2/2}$ and Mills' ratio,

$$\begin{aligned} I_2 &\le \mathbb{E}_0 X^2 \mathbf{1}\{|X| > a_n\} = 2 \int_{a_n}^\infty x^2 \phi(x) dx \\ &= 2[-x\phi(x)]_{a_n}^\infty + \int_{a_n}^\infty \phi(x) dx \le e^{-a_n^2/2} (2a_n + 1). \end{aligned}$$

Plugging the expression for a_n into the r.h.s. shows that $I_2 \lesssim (p_n/n) \log(n/p_n)$ as well and this finally gives $\mathbb{E}_0 (X m_X)^2 \lesssim (p_n/n) \log(n/p_n)$. □

Lemma 2.9. *Work under Conditions 2 and 3. Then,*

$$\sum_{i:\theta_i=0}^{n} \mathbb{E}_0 \operatorname{var}(\theta_i \mid X_i) \lesssim p_n \log(n/p_n).$$

Proof. Let $a_n = \sqrt{2\log(n/p_n)}$. It is enough to show that $\mathbb{E}_0 \operatorname{var}(\theta \mid X) \lesssim p_n \log(n/p_n)/n$. To prove this, we need to treat the cases that $|X|$ is larger/smaller than a_n, separately. To bound the variance, we use (2.13), that is $\operatorname{var}(\theta \mid X) \le m_x + x^2 m_x \le 1 + x^2$.

Case $|X| > a_n$: Using the identity $d/dx[x\phi(x)] = \phi(x) - x^2\phi(x)$,

$$\begin{aligned}\mathbb{E}_0 \operatorname{var}(\theta \mid X)\mathbf{1}_{\{|X|>a_n\}} &\le 2\int_{a_n}^{\infty}(1+x^2)\phi(x)dx = 2\Phi^c(a_n) + 2\int_{a_n}^{\infty} x^2\phi(x)dx \\ &= 4\Phi^c(a_n) + 2[-x\phi(x)]_{a_n}^{\infty} \le 4\phi(a_n) + 2a_n\phi(a_n). \end{aligned} \tag{2.20}$$

Using the expression for a_n shows that this can be bounded by $(p_n/n)\sqrt{\log(n/p_n)}$.

Case $|X| \le a_n$: Notice that the variance bound implies $\operatorname{var}(\theta \mid X) \le m_x\mathbf{1}\{|x| \le 1\} + 2x^2 m_x$. Below, we estimate $\mathbb{E}_0 m_X \mathbf{1}\{|X| \le 1\}$ and $\mathbb{E}_0 X^2 m_X \mathbf{1}\{|X| \le a_n\}$. For the first term, using (2.19),

$$\mathbb{E}_0 m_X \mathbf{1}\{|X| \le 1\} \lesssim \int_{-1}^{1}(s_n e^{x^2/4} + q_n e^{x^2/2})\phi(x)dx \le 4s_n. \tag{2.21}$$

For the second term $\mathbb{E}_0 X^2 m_X \mathbf{1}\{|X| \le a_n\}$, we use the second inequality in (2.19) and find

$$\begin{aligned}\mathbb{E}_0 X^2 m_X \mathbf{1}\{|X| \le a_n\} &\lesssim s_n \int_{-a_n}^{a_n} x^2 e^{\frac{x^2}{4}}\phi(x)dx \\ &\quad + \int_{-a_n}^{a_n}\int_{1}^{\infty} \frac{u\pi(u)}{(1+u)^{3/2}} x^2 e^{-\frac{x^2}{2+2u}} dudx.\end{aligned}$$

The first integral is bounded by a constant and for the second integral, we use Fubini's theorem, substitute $y = x/\sqrt{1+u}$, and use Condition 3

$$\begin{aligned}\int_{-a_n}^{a_n}\int_{1}^{\infty} \frac{u\pi(u)}{(1+u)^{3/2}} x^2 e^{-\frac{x^2}{2+2u}} dudx &= \int_{1}^{\infty} u\pi(u) \int_{-a_n/\sqrt{1+u}}^{a_n/\sqrt{1+u}} y^2 e^{-\frac{y^2}{2}} dydu \\ &\le \int_{1}^{\infty} u\pi(u)\left[\left(\frac{a_n}{\sqrt{1+u}}\right)^3 \wedge \sqrt{2\pi}\right] du \\ &\le 2^{3/2} C s_n.\end{aligned}$$

Together with (2.21) this shows that $\mathbb{E}_0 \operatorname{var}(\theta \mid X)\mathbf{1}\{|X| \le a_n\} \lesssim s_n$. Since in both cases the upper bound is of order $(p_n/n)\log(n/p_n)$ the result follows. □

3

Adaptive inference and uncertainty quantification for the horseshoe

hyperprior

Abstract

We investigate the frequentist properties of Bayesian procedures for estimation and uncertainty quantification based on the horseshoe prior. We consider the sparse multivariate mean model and consider both the hierarchical Bayes method of putting a prior on the unknown sparsity level and the empirical Bayes method with the sparsity level estimated by maximum marginal likelihood. We show that both Bayesian techniques lead to rate-adaptive optimal posterior contraction. We also investigate the frequentist coverage of Bayesian credible sets resulting from the horseshoe prior, both when the sparsity level is set by an oracle and when it is set by hierarchical or empirical Bayes. We show that credible balls and marginal credible intervals have good frequentist coverage and optimal size if the sparsity level of the prior is set correctly. By general theory honest confidence sets cannot adapt in size to an unknown sparsity level. Accordingly the hierarchical and empirical Bayes credible sets based on the horseshoe prior are not honest over the full parameter space. We show that this is due to over-shrinkage for certain parameters and characterise the set of parameters for which credible balls and marginal credible intervals do give correct uncertainty quantification. In particular we show that the fraction of false discoveries by the marginal Bayesian procedure is controlled

This chapter has been submitted as: S. van der Pas, B. Szabó and A. van der Vaart. How many needles in the haystack? Adaptive inference and uncertainty quantification for the horseshoe. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

by a correct choice of cut-off.

3.1 Introduction

The rise of big datasets with few signals, such as gene expression data and astronomical images, has given an impulse to the study of sparse models. The sequence model, or sparse normal means problem, is well studied. In this model, a random vector $Y^n = (Y_1, \ldots, Y_n)$ with values in $\mathbb{R}^n$ is observed, and each single observation Y_i is the sum of a fixed mean and standard normal noise:

$$Y_i = \theta_{0,i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where the ε_i are independent standard normal variables. We perform inference on the mean vector $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,n})$, and assume it to be sparse in the nearly black sense, meaning that all except an unknown number $p_n = \sum_{i=1}^n \mathbf{1}\{\theta_{0,i} \neq 0\}$ of the means are zero. We assume that p_n increases with n, but not as fast as n: $p_n \to \infty$ and $p_n/n \to 0$ as n tends to infinity.

Many methods to recover θ_0 have been suggested. Those most directly related to this work are Bhadra et al. (2015); Bhattacharya et al. (2014); Caron and Doucet (2008); Castillo et al. (2015); Castillo and Van der Vaart (2012); Ghosh and Chakrabarti (2015); Griffin and Brown (2010); Jiang and Zhang (2009); Johnson and Rossell (2010); Johnstone and Silverman (2004); Ročková (2015); Tibshirani (1996). In the present paper we study the Bayesian method based on the *horseshoe prior* (Carvalho et al., 2009, 2010; Polson and Scott, 2012*a*,*b*; Scott, 2011). Under this prior the coordinates $\theta_1, \ldots, \theta_n$ are an i.i.d. sample from a scale mixture of normals with a half-Cauchy prior on the variance, as follows. Given a "global hyperparameter" τ,

$$\begin{aligned} \theta_i \mid \lambda_i, \tau &\sim \mathcal{N}(0, \lambda_i^2 \tau^2), \\ \lambda_i &\sim C^+(0,1), \qquad i = 1, \ldots, n. \end{aligned} \tag{3.2}$$

In the Bayesian model the observations Y_i follow (3.1) with θ_0 taken equal to θ. The *posterior distribution* is then as usual obtained as the conditional distribution of θ given Y^n. For a given value of τ, possibly determined by an empirical Bayes method, aspects of the posterior distribution of θ, such as its mean and variance, can be computed with the help of analytic formulas and numerical integration (Van der Pas et al., 2014; Polson and Scott, 2012*a*,*b*). It is also possible to equip τ with a hyperprior, and follow a hierarchical, full Bayes approach. Several MCMC samplers and a software package are available for computation of the posterior distribution (Gramacy, 2014; Makalic and Schmidt, 2015; Scott, 2010).

The horseshoe posterior has performed well in simulations (Armagan et al., 2013; Bhattacharya et al., 2014; Carvalho et al., 2009, 2010; Polson and Scott, 2010, 2012*a*). Theoretical investigation in Van der Pas et al. (2014) shows that the parameter τ can, up to a logarithmic factor, be interpreted as the fraction of nonzero parameters θ_i. In particular, if τ is chosen to be at most of the order $(p_n/n)\sqrt{\log n/p_n}$, then the horseshoe posterior contracts to the true parameter at the (near) minimax rate of recovery for quadratic loss over sparse

models (Van der Pas et al., 2014). While motivated by these good properties, we also believe that the results obtained give insight in the performance of Bayesian procedures for sparsity in general.

In the present paper we make four novel contributions. First and second we establish the contraction rates of the posterior distributions of θ in the hierarchical, full Bayes case and in the general empirical Bayes case. Third we study the particular empirical Bayes method of estimating τ by the method of maximum Bayesian marginal likelihood. Fourth we study the capability of the posterior distribution for uncertainty quantification, in both the hierarchical and empirical Bayes cases.

As the parameter τ can be viewed as measuring sparsity, the first two contributions are both focused on adaptation to the number p_n of nonzero means, which is unlikely to be known in practice. The hierarchical and empirical Bayes methods studied here are shown to have similar performance, both in theory and in a small simulation study, and appear to outperform the ad-hoc estimator introduced in Van der Pas et al. (2014). The horseshoe posterior attains similar contraction rates as the spike-and-slab priors, as obtained in Castillo et al. (2015); Castillo and Van der Vaart (2012); Johnstone and Silverman (2004), and two-component mixtures, as in Ročková (2015). We obtain these results under general conditions on the hyperprior on τ, and for general empirical Bayes methods.

The conditions for the empirical Bayes method are met in particular by the maximum marginal likelihood estimator (MMLE). This is the maximum likelihood estimator of τ under the assumption that the "prior" (3.2) is part of the data-generating model, leaving only τ as a parameter. The MMLE is a natural estimator and is easy to compute. It turns out that the "MMLE plug-in posterior distribution" closely mimics the hierarchical Bayes posterior distribution. Besides practical benefit, this correspondence provides a theoretical tool to analyze the hierarchical Bayes method, which need not rely on testing arguments (as in Ghosal et al. (2000, 2008); Van der Vaart and Van Zanten (2009)).

In the Bayesian framework the spread of the posterior distribution over the parameter space is used as an indication of the error in estimation. For instance, a set of prescribed posterior probability around the center of the posterior distribution (a credible set) is often used in the same way as a confidence region for the parameter. A main contribution of the present paper is to investigate this practice for the horseshoe posterior distribution, in its dependence on the true signal θ_0. Besides for credible balls we also study this for credible intervals based on the marginal posterior distributions.

It follows from general results of Li (1989); Nickl and Van de Geer (2013); Robins and Van der Vaart (2006) that honest uncertainty quantification is irreconcilable with adaptation to sparsity. Here *honesty* of confidence sets $\hat{C}_n = \hat{C}_n(Y^n)$ relative to a parameter space $\tilde{\Theta} \subset \mathbb{R}^n$ means that

$$\liminf_{n\to\infty} \inf_{\theta_0\in\tilde{\Theta}} P_{\theta_0}(\theta_0 \in \hat{C}_n) \geq 1-\alpha,$$

for some prescribed confidence level $1-\alpha$. Furthermore, *adaptation* to a partition $\tilde{\Theta} = \cup_{p\in P}\Theta_p$ of the parameter space into submodels Θ_p indexed by a hyperparameter $p \in P$, means that, for every $p \in P$ and for $r_{n,p}$ the (near) minimax rate of estimation relative to Θ_p,

$$\liminf_{n\to\infty} \inf_{\theta_0\in\Theta_p} P_{\theta_0}(\mathrm{diam}(\hat{C}_n) \leq r_{n,p}) = 1.$$

This second property ensures that the good coverage is not achieved by taking conservative, overly large confidence sets, but that these sets have "optimal" diameter. In our present situation we may choose the models Θ_p equal to nearly black bodies with p nonzero coordinates, in which case $r_{n,p}^2 \asymp p\log(n/p)$, if $p \ll n$. Now it is shown in Li (1989) that confidence regions that are honest over all parameters in $\tilde{\Theta} = \mathbb{R}^n$ cannot be of square diameter smaller than $n^{1/2}$, which can be (much) bigger than $p\log(n/p)$, if $p \ll n$. Similar restrictions are valid for honesty over subsets of $\mathbb{R}^n$, as follows from testing arguments (see the appendix in Robins and Van der Vaart (2006)). Specifically, in Nickl and Van de Geer (2013) it is shown that confidence regions that adapt in size to nearly black bodies of two different dimensions $p_{n,1} \ll p_{n,2}$ cannot be honest over the union of these two bodies, but only over the union of the smallest body and the vectors in the bigger body that are at some distance from the smaller body. As both the full Bayes and empirical Bayes horseshoe posteriors contract at the near minimax rate $r_{n,p}$, adaptively over every nearly black body, it follows that their credible balls cannot be honest in the full parameter space.

In Bayesian practice credible balls are nevertheless used as if they were confidence sets. A main contribution of the present paper is to investigate for which parameters θ_0 this practice is justified. We characterise the parameters for which the credible sets of the horseshoe posterior distribution give good frequentist coverage, and the ones for which they do not. We investigate this both for the empirical and hierarchical Bayes approaches, both when τ is set deterministically, and in adaptive settings where the number of nonzero means is unknown. In the case of deterministically chosen τ, uncertainty quantification is essentially correct provided τ is chosen not smaller than $(p_n/n)\sqrt{\log n/p_n}$. For the more interesting full and empirical Bayes approaches, the correctness depends on the sizes of the nonzero coordinates in θ_0. If a fraction of the nonzero coordinates is detectable, meaning that they exceed the "threshold" $\sqrt{2\log(n/p)}$, then uncertainty quantification by a credible ball is correct up to a multiplicative factor in the radius. More generally, this is true if the sum of squares of the non-detectable nonzero coordinates is suitably dominated, as in Belitser and Nurushev (2015).

Uncertainty quantification for single coordinates $\theta_{0,i}$ by marginal credible intervals is quite natural. Credible intervals can be easily visualised by plotting them versus the index (cf. Figure 3.2). They may also be used as a testing device, for instance by declaring coordinates i for which the credible interval does not contain 0 to be *discoveries*. We show that the validity of these intervals depends on the value of the true coordinate. On the positive side we show that marginal credible intervals for coordinates $\theta_{0,i}$ that are either close to zero or above the detection boundary are essentially correct. In particular, the fraction of false discoveries (referring to zero $\theta_{0,i}$ that are declared nonzero) can be controlled by slightly enlarging the length of the intervals. On the negative side the horseshoe posteriors shrink intervals for intermediate values too much to zero for good frequentist coverage. Different from the case of credible balls, these conclusions are hardly affected by whether the sparseness level τ is set by an oracle or adaptively, based on the data.

We conclude that the uncertainty quantification given by the horseshoe posterior distribution is "honest" only conditionally on certain prior assumptions on the parameters. In contrast, interesting recent work within the context of the sparse linear regression model is directed at obtaining confidence sets that are honest in the full parameter set (Van de Geer et al., 2014; Liu and Yu, 2013; Zhang and Zhang, 2014). The resulting methodology, appro-

priately referred to as "de-sparsification", might in our present very special case of the regression model reduce to confidence sets for θ_0 based on the trivial pivot $Y^n - \theta_0$, or functions thereof, such as marginals. These confidence sets would have uniformly correct coverage, but be very wide, and not employ the presumed sparsity of the parameter. This seems a high price to pay; sacrificing some coverage so as to retain some shrinkage may not be unreasonable. Our contribution here is to investigate in what way the horseshoe prior makes this trade-off.

Uncertainty quantification in the case of the sparse normal mean model was addressed also in the recent paper by Belitser and Nurushev (2015). These authors consider a mixed Bayesian-frequentist procedure, which leads to a mixture over sets $I \subset \{1, 2, \ldots, n\}$ of projection estimators $(Y_i \mathbf{1}_{i \in I})$, where the weights over I have a Bayesian interpretation and each projection estimator comes with a distribution. Treating this as a posterior distribution, the authors obtain credible balls for the parameter, which they show to be honest over parameter vectors θ_0 that satisfy an "excessive-bias restriction". This interesting procedure has similar properties as the horseshoe posterior distribution studied in the present paper. While initially we had derived our results under a stronger "self-similarity" condition, we present here the results under a slight weakening of the "excessive-bias restriction" introduced in Belitser and Nurushev (2015).

The performance of adaptive Bayesian methods for uncertainty quantification for the estimation of functions has been previously considered in Castillo and Nickl (2014); Ray (2014); Serra and Krivobokova (2014); Szabó et al. (2015*a*); Szabó et al. (2015*b*), Belitser (2014); Sniekers and Van der Vaart (2015*a*,*b*,*c*). These papers focus on adaptation to functions of varying regularity. This runs into similar problems of honesty of credible sets, but the ordering by regularity sets the results apart from the adaptation to sparsity in the present paper.

The paper is organized as follows. We first introduce the MMLE in Section 3.2. Next we present contraction rates in Section 3.3, for general empirical and hierarchical Bayes approaches, and specifically for the MMLE. Coverage of credible balls and marginal credible intervals, again for general empirical and hierarchical Bayes approaches, and the MMLE in particular, are stated in Section 3.4. We illustrate the coverage properties of the marginal credible sets computed by empirical and hierarchical Bayes methods in a simulation study in Section 3.5. We conclude with Section 3.6, containing all proofs not given in the main text.

3.1.1 Notation

We use $\Pi(\cdot \mid Y^n, \tau)$ for the posterior distribution of θ relative to the prior (3.2) given fixed τ, and $\Pi(\cdot \mid Y^n)$ for the posterior distribution in the hierarchical setup where τ has received a prior. The empirical Bayes "plug-in posterior" is the first object with a data-based variable $\hat{\tau}$ substituted for τ. In order to stress that this does not entail conditioning on $\hat{\tau}$, we also write $\Pi_\tau(\cdot \mid Y^n)$ for $\Pi(\cdot \mid Y^n, \tau)$, and then $\Pi_{\hat{\tau}}(\cdot \mid Y^n)$ is the empirical Bayes (or plug-in) posterior distribution.

The density of the standard normal distribution is denoted by φ. Furthermore, $\ell_0[p] = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \mathbf{1}\{\theta_i \neq 0\} \leq p\}$ denotes the class of nearly black vectors, and we abbreviate

$$\zeta_\tau = \sqrt{2\log(1/\tau)}, \qquad \tau_n(p) = (p/n)\sqrt{\log(n/p)}, \qquad \tau_n = \tau_n(p_n).$$

3.2 Maximum marginal likelihood estimator

In this Section we define the MMLE and compare it to a naive empirical Bayes estimator previously suggested in Van der Pas et al. (2014). In Sections 3.3.1 and 3.4.2 we show that the MMLE is close to the "optimal" value $\tau_n(p_n) = (p_n/n)\sqrt{\log(n/p_n)}$ with high probability, leads to posterior contraction at the near-minimax rate, and yields adaptive confidence sets for selected parameters.

The marginal prior density of a parameter θ_i in the model (3.2) is given by

$$g_\tau(\theta) = \int_0^\infty \varphi\left(\frac{\theta}{\lambda\tau}\right)\frac{1}{\lambda\tau}\frac{2}{\pi(1+\lambda^2)}\,d\lambda. \tag{3.3}$$

In the Bayesian model the observations Y_i are distributed according to the convolution of this density and the standard normal density. The MMLE is the maximum likelihood estimator of τ in this latter model, given by

$$\widehat{\tau}_M = \underset{\tau\in[1/n,1]}{\operatorname{argmax}} \prod_{i=1}^n \int_{-\infty}^\infty \varphi(y_i-\theta)g_\tau(\theta)\,d\theta. \tag{3.4}$$

The restriction of the MMLE to the interval $[1/n, 1]$ can be motivated by the interpretation of τ as the level of sparsity, as in Van der Pas et al. (2014), which makes the interval correspond to assuming that at least one and at most all parameters are nonzero. The lower bound of $1/n$ has the additional advantage of preventing computational issues that arise when τ is very small (Datta and Ghosh (2013); Van der Pas et al. (2014)). We found the observation in Datta and Ghosh (2013) that an empirical Bayes approach cannot replace a hierarchical Bayes one, because the estimate of τ tends to be too small, too general. In both our theoretical study as in our simulation results the restriction that the MMLE be at least $1/n$ prevents a collapse to zero. Our simulations, presented in Section 3.5, also give no reason to believe that the hierarchical Bayes method is inherently better than empirical Bayes. Indeed, they behave very similarly (depending on the prior on τ).

An interpretation of τ as the fraction of nonzero coordinates motivates another estimator (Van der Pas et al. (2014)), which is based on a count of the number of observations that exceed the "universal threshold" $\sqrt{2\log n}$:

$$\widehat{\tau}_S(c_1,c_2) = \max\left\{\frac{\sum_{i=1}^n \mathbf{1}\{|y_i|\geq\sqrt{c_1\log n}\}}{c_2 n}, \frac{1}{n}\right\}, \tag{3.5}$$

where c_1 and c_2 are positive constants. If $c_2 > 1$ and ($c_1 > 2$ or $c_1 = 2$ and $p_n \gtrsim \log n$), then the plug-in posterior distribution with the *simple estimator* $\widehat{\tau}_S(c_1,c_2)$ contracts at the near square minimax rate $p_n \log n$ (see Section 1.4 in Chapter 1). This also follows from Theorem 3.2 in the present paper, as $\widehat{\tau}_S(c_1,c_2)$ satisfies Condition 4 below. On the other hand, this estimator fails to meet Condition 8 and hence our results on coverage do not apply to it. It appears that the simple estimator tends to be "too small".

This is corroborated by the numerical study presented in Figure 3.1. The figure shows approximations to the expected values of $\widehat{\tau}_S$ and $\widehat{\tau}_M$ when θ_0 is a vector of length $n = 100$, with p_n coordinates drawn from a $\mathcal{N}(A,1)$ distribution, with $A \in \{1,4,7\}$, and the remaining coordinates drawn from a $\mathcal{N}(0,1/4)$ distribution. For this sample size the "universal

threshold" $\sqrt{2\log n}$ is approximately 3, and thus signals with $A = 1$ should be difficult to detect, whereas those with $A = 7$ should be easy; those with $A = 4$ represent a boundary case.

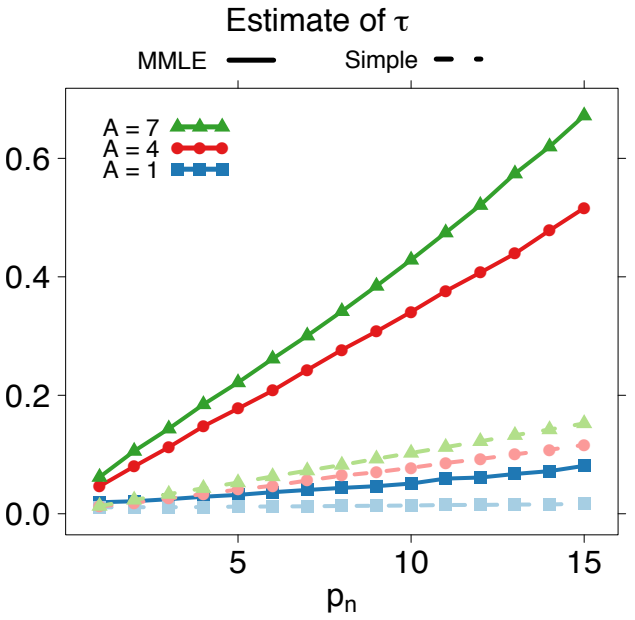

Figure 3.1: Approximate expected values of the MMLE (3.4) (solid) and the simple estimator (3.5) with $c_1 = 2$ and $c_2 = 1$ (dotted) when p_n (horizontal axis) out of $n = 100$ parameters are drawn from a $\mathcal{N}(A, 1)$ distribution, and the remaining $(n - p_n)$ parameters from a $\mathcal{N}(0, 1/4)$ distribution. The study was conducted with $A = 1$ (■), $A = 4$ (●) and $A = 7$ (▲). The results as shown are the averages over $N = 1000$ replications.

The figure shows that in all cases the MMLE (3.4) yields larger estimates of τ than the simple estimator (3.5), and thus leads to less shrinkage. This is expected in light of the results in the following section, which show that the MMLE is of order $\tau_n(p_n)$, whereas the simple estimator is capped at p_n/n. Both estimators appear to be linear in the number of nonzero coordinates of θ_0, with different slopes. When the signals are below the universal threshold, then the simple estimator is unlikely to detect any of them, whereas the MMLE may still pick up some of the signals. We study the consequences of this for the mean square errors and credible sets in Section 3.5.

3.3 Contraction rates

In this section we establish the rate of contraction of both the empirical Bayes and full Bayes posterior distributions. The *empirical Bayes posterior* is found by replacing τ in the

posterior distribution $\Pi(\cdot \mid Y^n, \tau)$ of θ relative to the prior (3.2) with a given τ by a data-based estimator $\widehat{\tau}$; we denote this by $\Pi_{\widehat{\tau}}(\cdot \mid Y^n)$. The *full Bayes posterior* $\Pi(\cdot \mid Y^n)$ is the ordinary posterior distribution of θ in the model where τ is also equipped with a prior and (3.2) is interpreted as the conditional prior of θ given τ.

The rate of contraction refers to properties of these posterior distributions when the vector Y^n follows a normal distribution on $\mathbb{R}^n$ with mean θ_0 and covariance the identity. We give general conditions on the empirical Bayes estimator $\hat{\tau}_n$ and the hyperprior on τ that ensure that the square posterior rate of contraction to θ_0 of the resulting posterior distributions is the near minimax rate $p_n \log n$ for estimation of θ_0 relative to the Euclidean norm. We also show that these conditions are met by the MMLE and natural hyperpriors on τ.

The minimax rate, the usual criterion for point estimators, has proven to be a useful benchmark for the speed of contraction of posterior distributions as well. The posterior cannot contract faster to the truth than at the minimax rate (Ghosal et al., 2000). The square minimax ℓ_2-rate for the sparse normal means problem is $p_n \log(n/p_n)$ (Donoho et al., 1992). This is slightly faster (i.e. smaller) than $p_n \log n$, but equivalent if the true parameter vector is not very sparse (if $p_n \leq n^\alpha$, for some $\alpha < 1$, then $(1-\alpha)p_n \log n \leq p_n \log(n/p_n) \leq p_n \log n$). For adaptive procedures, where the number of nonzero means p_n is unknown, results are usually given in terms of the "near-minimax rate" $p_n \log n$, for example for the spike-and-slab Lasso (Ročková, 2015), the Lasso (Bickel et al., 2009), and the horseshoe (Van der Pas et al., 2014).

3.3.1 Empirical Bayes

The empirical Bayes posterior distribution achieves the near-minimax contraction rate provided that the estimator $\widehat{\tau}_n$ of τ satisfies the following condition. Let $\tau_n(p) = (p/n)\sqrt{\log(n/p)}$.

Condition 4. There exists a constant $C > 0$ such that $\widehat{\tau}_n \in [1/n, C\tau_n(p_n)]$, with P_{θ_0}-probability tending to one, uniformly in $\theta_0 \in \ell_0[p_n]$.

This condition is weaker than the condition given in Van der Pas et al. (2014) for ℓ_2-adaptation of the empirical Bayes posterior mean, which requires asymptotic concentration of $\hat{\tau}_n$ on the same interval $[1/n, C\tau_n(p_n)]$ but at a rate. In Van der Pas et al. (2014) a plug-in value for τ of order $\tau_n(p_n)$ was found to be the largest value of τ for which the posterior distribution contracts at the minimax rate, and has variance of the same order. Condition 4 can be interpreted as ensuring that $\widehat{\tau}_n$ is of at most this "optimal" order. The lower bound can be interpreted as assuming that there is at least one nonzero mean, which is reasonable in light of the assumption $p_n \to \infty$. In addition, it prevents computational issues, as discussed in Section 3.2.

A main result of the present paper is that the MMLE satisfies Condition 4.

Theorem 3.1. The MMLE (3.4) satisfies Condition 4.

Proof. See Appendix 3.6.1. □

A second main result is that under Condition 4 the posterior contracts at the near-minimax rate.

Theorem 3.2. For any estimator $\widehat{\tau}_n$ of τ that satisfies Condition 4, the empirical Bayes posterior distribution contracts around the true parameter at the near-minimax rate: for any $M_n \to \infty$ and $p_n \to \infty$,

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_{\widehat{\tau}_n}\Big(\theta : \|\theta_0 - \theta\|_2 \geq M_n\sqrt{p_n \log n} \,|\, Y^n\Big) \to 0.$$

In particular, this is true for $\widehat{\tau}_n$ equal to the MMLE.

Proof. See Appendix 3.6.3. □

3.3.2 Hierarchical Bayes

The full Bayes posterior distribution contracts at the near minimax rate whenever the prior density π_n on τ satisfies the following two conditions.

Condition 5. The prior density π_n is supported inside $[1/n, 1]$.

Condition 6. Let $t_n = C_u \pi^{3/2} \tau_n(p_n)$, with the constant C_u as in Lemma 3.28(i). The prior density π_n satisfies

$$\int_{t_n/2}^{t_n} \pi_n(\tau)\, d\tau \gtrsim e^{-cp_n}, \quad \text{for some } c > C_u/10.$$

The restriction of the prior distribution to the interval $[1/n, 1]$ can be motivated by the same reasons as discussed under the definition of the MMLE in Section 3.2. In our simulations (also see Chapter 1) we have also noted that large values produced by for instance a sampler using a half-Cauchy prior, as in the original set-up proposed by Carvalho et al. (2010), were not beneficial to recovery.

As t_n is of the same order as $\tau_n(p_n)$, Condition 6 is similar to Condition 4 in the empirical Bayes case. It requires that there is sufficient prior mass around the "optimal" values of τ. The condition is satisfied by many prior densities, including the usual ones, except in the very sparse case that $p_n \lesssim \log n$, when it requires that π_n is unbounded near zero. For this situation we also introduce the following weaker condition, which is still good enough for a contraction rate with additional logarithmic factors.

Condition 7. For t_n as in Condition 6 the prior density π_n satisfies,

$$\int_{t_n/2}^{t_n} \pi_n(\tau)\, d\tau \gtrsim t_n.$$

Example 3.3. The Cauchy distribution on the positive reals, truncated to $[1/n, 1]$, has density $\pi_n(\tau) = (\arctan(1) - \arctan(1/n))^{-1}(1+\tau^2)^{-1}\mathbf{1}_{\tau \in [1/n,1]}$. This satisfies Condition 5, of course, and Condition 7. It also satisfies the stronger Condition 6 provided $t_n \geq e^{-cp_n}$, i.e. $p_n \geq C \log n$, for a sufficiently large C.

Example 3.4. For the uniform prior on $[1/n, 1]$, with density $\pi_n(\tau) = n/(n-1)\mathbf{1}_{\tau \in [1/n,1]}$, the same conclusions hold.

Example 3.5. For the prior with density $\pi_n(x) \propto 1/x$ on $[1/n, 1]$, Conditions 5 and 6 hold provided $p_n \gg \operatorname{loglog} n$.

The following lemma is a crucial ingredient of the derivation of the contraction rate. It shows that the posterior distribution of τ will concentrate its mass at most a constant multiple of t_n away from zero. We denote the posterior distribution of τ by the same general symbol $\Pi(\cdot \mid Y^n)$.

Lemma 3.6. If Conditions 5 and 6 hold, then

$$\inf_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi(\tau : \tau \leq 5t_n \mid Y^n) \to 1.$$

Furthermore, if only Conditions 5 and 7 hold, then a similar assertion is true, but with $5t_n$ replaced by $(\log n)t_n$.

Proof. See Appendix 3.6.3. □

We are ready to state the posterior contraction result for the full Bayes posterior.

Theorem 3.7. If the prior on τ satisfies Conditions 5 and 6, then the hierarchical Bayes posterior contracts to the true parameter at the near minimax rate: for any $M_n \to \infty$ and $p_n \to \infty$,

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi(\theta : \|\theta - \theta_0\|_2 \geq M_n \sqrt{p_n \log n} \mid Y^n) \to 0.$$

If the prior on τ satisfies only Conditions 5 and 7, then this is true with $\sqrt{p_n \log n}$ replaced by $\sqrt{p_n} \log n$.

Proof. Using the notation $r_n = \sqrt{p_n \log n}$, we can decompose the left side of the preceding display as

$$\begin{aligned} &\mathbb{E}_{\theta_0} \Big[\int_{\tau \leq 5t_n} + \int_{\tau > 5t_n} \Big] \Pi_\tau(\theta : \|\theta - \theta_0\|_2 \geq M_n r_n \mid Y^n)\, \pi(\tau \mid Y^n)\, d\tau \\ &\quad \leq \mathbb{E}_{\theta_0} \sup_{\tau \leq 5t_n} \Pi_\tau(\theta : \|\theta - \theta_0\|_2 \geq M_n r_n \mid Y^n) + \mathbb{E}_{\theta_0} \Pi(\tau : \tau > 5t_n \mid Y^n). \end{aligned}$$

The first term on the right tends to zero by Theorem 3.2, and the second by Lemma 3.6. □

3.4 Coverage

By their definition *credible sets* contain a fixed fraction, e.g. 95 %, of the posterior mass. The diameter of such sets will be at most of the order of the posterior contraction rate. The upper bounds on the contraction rates of the horseshoe posterior distributions given in Section 3.3 imply that these are narrow enough to be informative. However, these bounds do not guarantee that the credible sets will *cover* the truth. The latter is dependent on the *spread* of the posterior mass relative to its distance to the true parameter. For instance, the bulk of the posterior mass may be highly concentrated inside a ball of radius

the contraction rate, but within a narrow area of diameter much smaller than its distance to the true parameter.

In this section we study coverage first in the case of deterministic τ and next for the empirical and full Bayes posterior distributions. We consider both credible balls (for the full parameter vector $\theta_0 \in \mathbb{R}^n$ relative to the Euclidean distance) and credible intervals (for the individual coordinates $\theta_{0,i}$). The latter are based on the marginal posterior distributions of the coordinates θ_i.

In Section 3.4.1 we show that a (slightly enlarged) credible ball centered at the posterior mean covers the truth provided τ is chosen bigger than the "optimal" value $\tau_n(p_n)$. Furthermore, we show that the marginal credible intervals fall into three categories, dependent on τ. For coordinates $\theta_{0,i}$ with absolute value below a multiple of τ or above a multiple of ζ_τ the credible intervals will cover, in the sense that within both categories the fraction of correct intervals is arbitrarily close to 1. On the other hand, none of the intermediate coordinates $\theta_{0,i}$ are covered.

In Section 3.4.2 we consider the case that p_n is not known, and the posterior is adapted to the sparsity level by either the empirical or the full Bayes method. Here the potential problem for coverage of credible balls is the *over-shrinkage* of the posterior distributions, due to a too small value of the MMLE $\widehat{\tau}_M$ or concentration of the posterior distribution of τ too close to zero. We show that such over-shrinkage does not occur, and both empirical and hierarchical credible balls cover, if the true parameter θ_0 satisfies the "excessive-bias restriction", given below. Furthermore, we show that the results for deterministic marginal credible intervals extend to the adaptive situation for *any* true parameter θ_0, with slight modification of the boundaries between the three cases of small, intermediate and large coordinates.

3.4.1 Credible sets for deterministic τ

Given a deterministic hyperparameter τ, possibly depending on n and p_n, we consider a *credible ball* of the form

$$\hat{C}_n(L,\tau) = \Big\{\theta : \|\theta - \hat{\theta}(\tau)\|_2 \le L\hat{r}(\alpha,\tau)\Big\}, \tag{3.6}$$

where $\hat{\theta}(\tau) = \mathbb{E}(\theta \mid Y^n, \tau)$ is the posterior mean, L a positive constant, and for a given $\alpha \in (0,1)$ the number $\hat{r}(\alpha,\tau)$ is determined such that

$$\Pi\Big(\theta : \|\theta - \hat{\theta}(\tau)\|_2 \le \hat{r}(\alpha,\tau) \mid Y^n, \tau\Big) = 1-\alpha.$$

Thus $\hat{r}(\alpha,\tau)$ is the natural radius of a set of "Bayesian credible level" $1-\alpha$, and L is a constant, introduced to make up for a difference between credible and confidence levels, similarly as in Szabó et al. (2015*a*). (Unlike in the latter paper the radii $\hat{r}(\alpha,\tau)$ do depend on the observation Y^n, as indicated by the hat in the notation.)

The following lower bound for $\hat{r}(\alpha,\tau)$ in the case that $n\tau \to \infty$ is the key to the frequentist coverage. The assumption $n\tau/\zeta_\tau \to \infty$ is satisfied for τ of the order the "optimal" rate $\tau_n(p_n)$ provided $p_n \to \infty$ (as we assume).

Lemma 3.8. If $n\tau/\zeta_\tau \to \infty$, then with P_{θ_0}-probability tending to one,

$$\hat{r}(\alpha,\tau) \ge 0.5\sqrt{n\tau\zeta_\tau}.$$

Proof. See Section 3.6.4. □

Theorem 3.9. If $\tau \geq \tau_n$ and $\tau \to 0$ and $p_n \to \infty$ with $p_n = o(n)$, then, there exists a large enough $L > 0$ such that

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \ell_0[p_n]} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L,\tau)\Big) \geq 1 - \alpha.$$

Proof. The probability of the complement of the event in the display is equal to $P_{\theta_0}\Big(\|\theta_0 - \hat{\theta}(\tau)\|_2 > L\,\hat{r}(\alpha,\tau)\Big)$. In view of Lemma 3.8 this is bounded by $o(1)$ plus

$$P_{\theta_0}\Big(\|\theta_0 - \hat{\theta}(\tau)\|_2 > 0.5L\sqrt{n\tau\zeta_\tau}\Big) \lesssim \frac{\mathbb{E}_{\theta_0}\|\hat{\theta}(\tau) - \theta_0\|_2^2}{L^2 n\tau\zeta_\tau}.$$

By Theorem 1.2 of Chapter 1 (or see the proof of Theorem 3.2 below) the numerator on the right is bounded by a multiple of $p_n \log(1/\tau) + n\tau\sqrt{\log 1/\tau}$. By the assumption $\tau \geq \tau_n \geq 1/n$ the quotient is smaller than α for appropriately large choice of L. □

Marginal credible intervals can be constructed from the *marginal posterior distributions* $\Pi(\theta : \theta_i \in \cdot \mid Y^n, \tau)$. By the independence of the pairs (θ_i, Y_i) given τ, the ith marginal depends only on the ith observation Y_i. We consider intervals of the form

$$\hat{C}_{ni}(L,\tau) = \Big\{\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq L\hat{r}_i(\alpha,\tau)\Big\}, \tag{3.7}$$

where $\hat{\theta}_i(\tau) = \mathbb{E}(\theta_i \mid Y_i, \tau)$ is the marginal posterior mean, L a positive constant, and $\hat{r}_i(\alpha,\tau)$ is determined so that, for a given $0 < \alpha \leq 1/2$,

$$\Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq \hat{r}_i(\alpha,\tau) \mid Y_i, \tau\Big) = 1 - \alpha.$$

The coverage of these intervals depends crucially on the value of the true coordinate $\theta_{0,i}$. For given $\tau \to 0$, positive constants k_S, k_M, k_L and numbers $f_\tau \uparrow \infty$ as $\tau \to 0$, we distinguish three regions (small, medium and large) of signal parameters:

$$\begin{aligned} S &:= \Big\{1 \leq i \leq n : |\theta_{0,i}| \leq k_S\tau\Big\},\\ M &:= \Big\{1 \leq i \leq n : f_\tau\tau \leq |\theta_{0,i}| \leq k_M\zeta_\tau\Big\},\\ L &:= \Big\{1 \leq i \leq n : k_L\zeta_\tau \leq |\theta_{0,i}|\Big\}. \end{aligned}$$

The conditions on the constants and f_τ in the following theorem make that these three sets may not cover all coordinates $\theta_{0,i}$, but their boundaries are almost contiguous. The following theorem shows that the fractions of coordinates contained in S and in L that are covered by the credible intervals are close to 1, whereas no coordinate in M is covered. Inspection of the proof will show that the latter occurs, because the corresponding intervals are shrunk too much to zero. Since all zero coordinates are in the set S, an overall conclusion is then that the set of "discoveries", the coordinates whose credible set does not contain 0, contains only a small fraction of "false discoveries". (In our setting the usual "false discovery rate" is not a useful quantity, as the number of nonzero parameters is a vanishing fraction of the total set of coordinates by assumption. The quantities considered in the theorem seem more descriptive of the accuracy of the procedure.)

Let $|\cdot|$ denote the cardinality of a set.

Theorem 3.10. Suppose that $k_S > 0$, $k_M < 1$, $k_L > 1$, and $f_\tau \uparrow \infty$, as $\tau \to 0$. Then for $\tau \to 0$ and any sequence $\gamma_n \to c$ for some $0 \le c \le 1/2$, satisfying $\zeta_{\gamma_n} \ll \zeta_\tau$,

$$P_{\theta_0}\Big(\frac{1}{|S|}|\{i \in S : \theta_{0,i} \in \hat{C}_{ni}(L_S, \tau)\}| \ge 1 - \gamma_n\Big) \to 1, \tag{3.8}$$

$$P_{\theta_0}\Big(\theta_{0,i} \notin \hat{C}_{ni}(L, \tau)\Big) \to 1, \quad \text{for any } L > 0 \text{ and } i \in M, \tag{3.9}$$

$$P_{\theta_0}\Big(\frac{1}{|L|}|\{i \in L : \theta_{0,i} \in \hat{C}_{ni}(L_L, \tau)\}| \ge 1 - \gamma_n\Big) \to 1, \tag{3.10}$$

where $L_S = (2.1/z_\alpha)\Big[k_S + (2/\gamma_n)\zeta_{\gamma_n/2}\Big]$ and $L_L = (1.1/z_\alpha)\zeta_{\gamma_n/2}$.

Proof. See Section 3.6.4. □

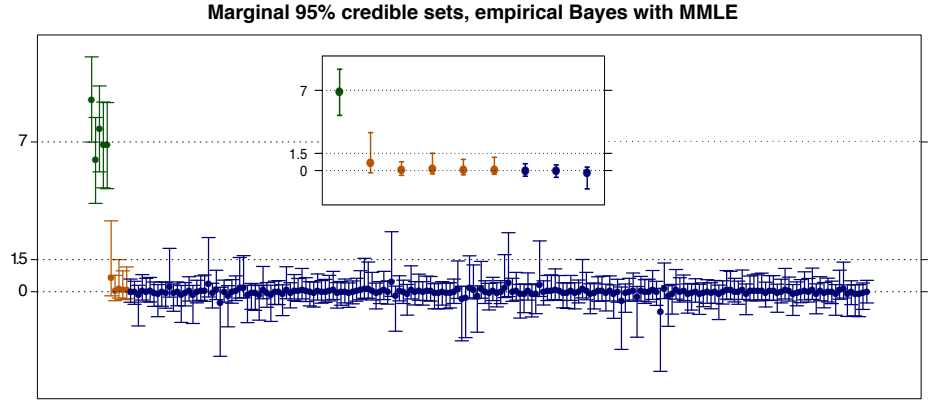

Figure 3.2: 95% marginal credible intervals based on the MMLE empirical Bayes method, for a single observation Y^n of length $n = 200$ with $p_n = 10$ nonzero parameters, the first 5 (from the left) being 7 (green), the next 5 equal to 1.5 (orange); the remaining 190 parameters are coded (blue). The inserted plot zooms in on credible intervals 5 to 13, thus showing one large mean and all intermediate means.

Figure 3.2 illustrates Theorem 3.10 by showing the marginal credible sets for just a single draw of the observation, in a setting with $n = 200$, and $p_n = 10$ nonzero coordinates. The value τ was chosen equal to the MMLE, which realised as approximately 0.11. The means were taken equal to 7, 1.5 or 0, corresponding to the three regions L, M, S listed in the theorem ($\sqrt{2 \log n} \approx 3.3$). All the large means (equal to 7) were covered; only 2 out of 5 of the medium means (equal to 1.5) were covered; and all small (zero) means were covered, in agreement with Theorem 3.10. It may be noted that intervals for zero coordinates are not necessarily narrow.

3.4.2 Adaptive credible sets

We now turn to credible sets in the more realistic scenario that the sparsity parameter p_n is not available. We investigate both the empirical Bayes and the hierarchical Bayes credible sets, and consider both balls and marginal intervals.

In the empirical Bayes approach we define a credible set by plugging in an estimator $\widehat{\tau}_n$ of τ into the non-adaptive credible ball $\hat{C}_n(L,\tau)$ given in (3.6):

$$\hat{C}_n(L,\widehat{\tau}_n) = \Big\{\theta : \|\theta - \hat{\theta}(\widehat{\tau}_n)\|_2 \leq L\hat{r}(\alpha,\widehat{\tau}_n)\Big\}. \tag{3.11}$$

In the hierarchical Bayes case we use a ball around the full posterior mean $\hat{\theta} = \int \theta\,\Pi(d\theta \mid Y^n)$, given by

$$\hat{C}_n(L) = \Big\{\theta : \|\theta - \hat{\theta}\|_2 \leq L\hat{r}(\alpha)\Big\}, \tag{3.12}$$

where L is a positive constant and $\hat{r}(\alpha)$ is defined from the full posterior distribution by

$$\Pi\Big(\theta : \|\theta - \hat{\theta}\|_2 \leq \hat{r}(\alpha) \mid Y^n\Big) = 1 - \alpha.$$

The question is whether these Bayesian credible sets are appropriate for uncertainty quantification from a frequentist point of view.

Unfortunately, coverage can be guaranteed only for a selection of true parameters θ_0. The problem is that a data-based estimate of sparsity may lead to *over-shrinkage*, which makes the credible sets too small and close to zero. A simple condition preventing over-shrinkage is that a sufficient number of nonzero parameters $\theta_{0,i}$ is above the "detection boundary". It turns out that the correct threshold for detection is given by $\sqrt{2\log(n/p_n)}$. This leads to the following condition.

Assumption 1 (self-similarity)**.** A vector $\theta_0 \in \ell_0[p]$ is called *self-similar* if

$$\#\Big(i : |\theta_{0,i}| \geq A\sqrt{2\log(n/p)}\Big) \geq \frac{p}{C_s}. \tag{3.13}$$

The two constants C_s and A will be fixed to universal values, where necessarily $C_s \geq 1$ and it is required that $A > 1$.

The problem of over-shinkage is comparable to the problem of over-smoothing in the context of nonparametric density estimation or regression, due to the choice of a too large bandwidth or smoothness level. The preceding self-similarity condition plays the same role as the assumptions of "self-similarity" or "polished tail" used by Bull (2012); Giné and Nickl (2010); Nickl and Szabó (2014); Picard and Tribouley (2000); Sniekers and Van der Vaart (2015*c*); Szabó et al. (2015*a*) in their investigations of confidence sets in nonparametric density estimation and regression, or the "excessive-bias" restriction in Belitser (2014) employed in the context of Besov-regularity classes in the normal mean model.

The self-similarity condition is also reminiscent of the *beta-min condition* for the adaptive Lasso (Bühlmann and Van de Geer, 2011; Van de Geer et al., 2011), which imposes a lower bound on the nonzero signals in order to achieve consistent selection of the set of nonzero coordinates of θ_0. However, the present condition is different in spirit both by the

size of the cut-off and by requiring only that a fraction of the nonzero means is above the threshold.

For ensuring coverage of credible balls the condition can be weakened to the following more technical condition.

Assumption 2 (excessive-bias restriction)**.** A vector $\theta_0 \in \ell_0[p]$ satisfies the *excessive-bias restriction* for constants $A > 1$ and $C_s, C > 0$, if there exists an integer $q \geq 1$ with

$$\sum_{i:|\theta_{0,i}|<A\sqrt{2\log(n/q)}} \theta_{0,i}^2 \leq Cq\log(n/q), \qquad \#\Big(i : |\theta_{0,i}| \geq A\sqrt{2\log(n/q)}\Big) \geq \frac{q}{C_s}. \tag{3.14}$$

The set of all such vectors θ_0 (for fixed constants A, C_s, C) is denoted by $\Theta[p]$, and $\tilde{p} = \tilde{p}(\theta_0)$ denotes $\#\Big(i : |\theta_{0,i}| \geq A\sqrt{2\log(n/q)}\Big)$, for the smallest possible q.

If $\theta_0 \in \ell_0[p]$ is self-similar, then it satisfies the excessive-bias restriction with $q = p$, $C = 2A^2$ and the same constants A and C_s. This follows, because the sum in (3.14) is trivially bounded by $\#(i : \theta_{0,i} \neq 0)\, A^2 2\log(n/q)$.

In the following example we show that the excessive-bias restriction is also implied by a condition with the same name introduced in Belitser and Nurushev (2015). The latter condition motivated Assumption 2, which is more suited to our investigation of the horseshoe credible sets.

Example 3.11. For a given θ_0 and any subset $I \subset \{1, 2, \ldots, n\}$ let

$$G(I) = \sum_{i \in I^c} \theta_{0,i}^2 + 2A^2|I| \log \frac{ne}{|I|}.$$

In Belitser and Nurushev (2015) θ_0 is defined to satisfy the *excessive-bias restriction* if G takes its minimum at a nonempty set $\tilde{I}$ such that $G(\tilde{I}) \leq C|\tilde{I}|\log(ne/|\tilde{I}|)$.

We now show that in this case θ_0 also satisfies Assumption 2, with $q = |\tilde{I}|$. Let $\theta_{0,i}$ be a coordinate with $i \in \tilde{I}$ of minimal absolute value $|\theta_{0,i}| = \min\{|\theta_{0,j}| : j \in \tilde{I}\}$. From $G(\tilde{I}) \leq G(\tilde{I} - \{i\})$ we obtain that $\theta_{0,i}^2 \geq 2A^2|\tilde{I}|\log(ne/|\tilde{I}|) - 2A^2(|\tilde{I}| - 1)\log(ne/(|\tilde{I}| - 1)) \geq 2A^2\log(n/|\tilde{I}|)$, since the derivative of $x \mapsto x\log(ne/x)$ is $\log(n/x)$. Consequently, first $\#(j : \theta_{0,j}^2 \geq 2A^2\log(n/|\tilde{I}|)) \geq \#(j : \theta_{0,j}^2 \geq \theta_{0,i}^2) \geq |\tilde{I}|$, by the minimising property of $\theta_{0,i}$, verifying the second inequality in (3.14). Second $\{j : \theta_{0,j}^2 < 2A^2\log(n/q)\} \subset \{j : \theta_{0,j}^2 < \theta_{0,i}^2\} \subset \tilde{I}^c$, again by the minimising property of $\theta_{0,i}$. Thus the first inequality of (3.14) follows by the fact that $G(\tilde{I}) \leq C|\tilde{I}|\log(ne/|\tilde{I}|)$.

To obtain coverage in the empirical Bayes setting, we replace Condition 4 by the following.

Condition 8. The estimator $\widehat{\tau}_n$ satisfies, for a given sequence p_n and some constant $C > 1$, with $\tilde{p} = \tilde{p}(\theta_0)$,

$$\inf_{\theta_0 \in \Theta[p_n]} P_{\theta_0}\Big(C^{-1}\tau_n(\tilde{p}) \leq \widehat{\tau}_n \leq C\tau_n(\tilde{p})\Big) \to 1.$$

Although this condition may appear more restrictive than Condition 4, as it requires a lower bound on $\widehat{\tau}_n$ of order $\tau_n(\tilde{p})$ instead of $1/n$, Condition 8 may not be more stringent than Condition 4, because it only needs to hold for vectors θ_0 that meet the excessive-bias restriction.

Lemma 3.12. For $p_n \to \infty$ such that $p_n = o(n)$, the MMLE $\widehat{\tau}_n$ satisfies Condition 8.

Proof. See Section 3.6.1. □

Theorem 3.13. Let $\tilde{p}_n \le p_n$ be given sequences with $\tilde{p}_n \to \infty$ and $p_n = o(n)$. If the estimator $\widehat{\tau}_n$ of τ satisfies Condition 8, then for a sufficiently large constant L the empirical Bayes credible ball $\hat{C}_n(L, \hat{\tau}_n)$ has honest coverage and rate adaptive (oracle) size:

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \Theta[p_n], \tilde{p}(\theta_0) \ge \tilde{p}_n} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L, \hat{\tau}_n)\Big) \ge 1 - \alpha,$$
$$\inf_{\theta_0 \in \Theta[p_n]} P_{\theta_0}\Big(\operatorname{diam}\big(\hat{C}_n(L, \widehat{\tau}_n)\big) \lesssim \sqrt{\tilde{p}\log(n/\tilde{p})}\Big) \to 1.$$

In particular, these assertions are true for the MMLE. Furthermore, if $\tilde{p}_n \ge C \log n$ for a sufficiently large constant C, then the hierarchical Bayes method with $\tau \sim \pi_n$ for π_n probability densities on $[1/n, 1]$ that are bounded away from zero also yields adaptive and honest confidence sets: for sufficiently large L,

$$\liminf_{n\to\infty} \inf_{\theta_0 \in \Theta[p_n], \tilde{p}(\theta_0) \ge \tilde{p}_n} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L)\Big) \ge 1 - \alpha,$$
$$\inf_{\theta_0 \in \Theta[p_n], \tilde{p}(\theta_0) \ge \tilde{p}_n} P_{\theta_0}\Big(\operatorname{diam}\big(\hat{C}_n(L)\big) \lesssim \sqrt{\tilde{p}\log(n/\tilde{p})}\Big) \to 1.$$

Proof. See Section 3.6.5. □

It may be noted that for self-similar θ_0 the square diameter of the credible balls is of the order $p\log(n/p)$, improving on the square contraction rate $p\log n$ obtained in Theorem 3.2. For parameters satisfying the excessive-bias restriction, this may further improve to $\tilde{p}\log(n/\tilde{p})$.

Adaptive empirical Bayes marginal credible intervals are defined by plugging in an estimator $\widehat{\tau}_n$ for τ in the intervals $\hat{C}_{ni}(L, \tau)$ defined by (3.7) in Section 3.4.1. Similarly full Bayes credible intervals $\hat{C}_{ni}(L)$ are defined from the full Bayes marginal posterior distributions. The following theorem shows that these intervals mimic the behaviour of the intervals for deterministic τ given in Theorem 3.14. In contrast to the case for credible balls, for this result the excessive-bias restriction is not required.

For given positive constants k_S, k_M, k_L, and f_n the three regions (small, medium and large) of signal parameters are defined as, where $p_n = \#(i : \theta_{0,i} \ne 0)$,

$$S_a := \Big\{1 \le i \le n : |\theta_{0,i}| \le k_S/n\Big\},$$
$$M_a := \Big\{1 \le i \le n : f_n\tau_n(p_n) \le |\theta_{0,i}| \le k_M\sqrt{2\log(1/\tau_n(p_n))}\Big\},$$
$$L_a := \Big\{1 \le i \le n : k_L\sqrt{2\log n} \le |\theta_{0,i}|\Big\}.$$

Theorem 3.14. Suppose that $k_S > 0$, $k_M < 1$, $k_L > 1$, and $f_n \uparrow \infty$. If $\hat{\tau}_n$ satisfies Condition 4, then for any sequence $\gamma_n \to c$ for some $0 \le c \le 1/2$ such that $\zeta^2_{\gamma_n} \ll \log(1/\tau_n(p_n))$, we have that

$$P_{\theta_0}\Big(\frac{1}{|S_a|}|\{i \in S_a : \theta_{0,i} \in \hat{C}_{ni}(L_S, \hat{\tau}_n)\}| \ge 1 - \gamma_n\Big) \to 1, \tag{3.15}$$

$$P_{\theta_0}\Big(\theta_{0,i} \notin \hat{C}_{ni}(L, \hat{\tau}_n)) \to 1, \quad \text{for any } L > 0 \text{ and } i \in M_a, \tag{3.16}$$

$$P_{\theta_0}\Big(\frac{1}{|L_a|}|\{i \in L_a : \theta_{0,i} \in \hat{C}_{ni}(L_L, \hat{\tau}_n)\}|/l \geq 1 - \gamma_n\Big) \to 1, \tag{3.17}$$

with L_S and L_L given in Theorem 3.10. Under Conditions 5 and 6 and in addition $p_n \gtrsim \log n$ the same statements hold for the hierarchical Bayes marginal credible sets. This is also true under Conditions 5 and 7 if $f_n \gg \log n$, with different constants L_S and L_L.

Proof. See Section 3.6.5. □

Remark 3.15. Under the self-similarity assumption (3.13) the statements of Theorem 3.14 hold for the sets S, M and L given preceding Theorem 3.10 with $\tau = \tau_n(p_n)$.

3.5 Simulation study

We study the relative performances of the empirical Bayes and hierarchical Bayes approaches further through simulation studies, extending the simulation study in Chapter 1. We first consider the mean square error (MSE) for empirical Bayes combined with either (i) the simple estimator (with $c_1 = 2, c_2 = 1$) or (ii) the MMLE, and for hierarchical Bayes with either (iii) a Cauchy prior on τ, or (iv) a Cauchy prior truncated to $[1/n, 1]$ on τ. We then study the coverage and average lengths of the marginal credible intervals resulting from these four methods, as well as intervals based solely on the posterior mean and variance.

3.5.1 Mean square error

We created a ground truth θ_0 of length $n = 400$ with $p_n \in \{20, 200\}$, where each nonzero mean was fixed to $A \in \{1, 2, \ldots, 10\}$. We computed the posterior mean for each of the four procedures, and approximated the MSE by averaging over $N = 100$ iterations. The results are shown in Figure 3.3. In addition the figure shows the MSE separately for the nonzero and zero coordinates of θ_0, and the average value (of the posterior mean) of τ.

The shapes of the curves of the overall MSE for methods (i) and (iii) were discussed in Chapter 1. Values close to the threshold $\sqrt{2 \log n} \approx 3.5$ pose the most difficult problem, and hierarchical Bayes with a Cauchy prior performs better below the threshold, while empirical Bayes with the simple estimator performs better above, as the simple estimator is very close to p_n/n in those settings, whereas the values of τ resulting from hierarchical Bayes are much larger.

Three new features stand out in this comparison, with the MMLE and hierarchical Bayes with a truncated Cauchy added in, and the opportunity to study the zero and nonzero means separately. The first is that empirical Bayes with the MMLE and hierarchical Bayes with the Cauchy prior truncated to $[1/n, 1]$ behave very similarly, as was expected from our proofs, in which the comparison of the two methods is fruitfully explored.

Secondly, while in the most sparse setting ($p_n = 20$), full Bayes with the truncated and non-truncated Cauchy priors yield very similar results, as the mean value of τ does not come close to the 'maximum' of 1 in either approach, the truncated Cauchy (and the MMLE) offer an improvement over the non-truncated Cauchy in the less sparse ($p_n = 200$)

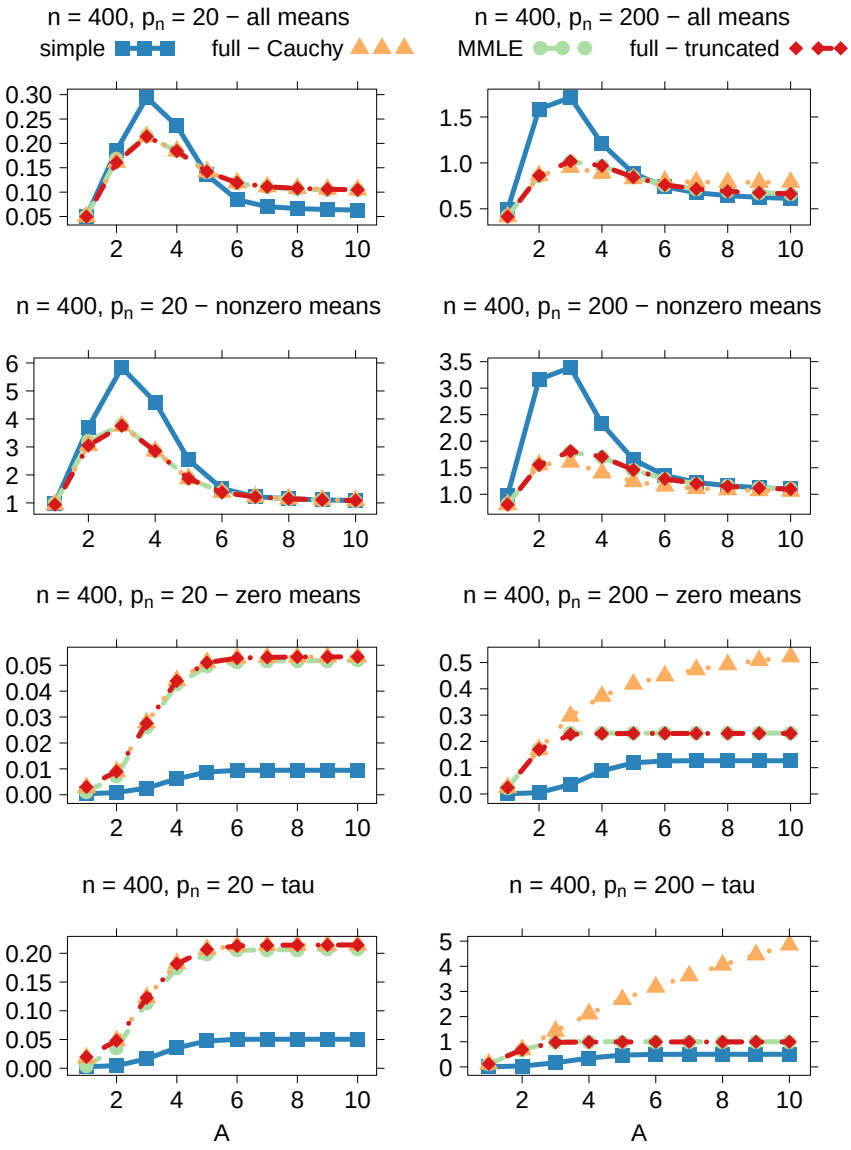

Figure 3.3: Mean square error (overall, for the nonzero coordinates, and for the zero coordinates) of the posterior mean corresponding to empirical Bayes with the simple estimator with $c_1 = 2, c_2 = 1$ (■) or the MMLE (●) and to hierarchical Bayes with a Cauchy prior on τ (▲) or a Cauchy prior truncated to $[1/n, 1]$ (♦). The bottom plot shows the average estimated value of τ (or the posterior mean in the case of the hierarchical Bayes approaches). The settings are $n = 400$ and $p_n = 20$ (left) and $p_n = 200$ (right); the results are approximations based on averaging over $N = 100$ samples for each value of A.

setting. The non-truncated Cauchy does lead to lower MSE on the nonzero means close to the threshold, but overestimates the zero means due to the large values of τ. With the MMLE and the truncated Cauchy, the restriction to $[1/n, 1]$ prevents the marginal posterior of τ from concentrating too far away from the 'optimal' values of order $\tau_n(p_n)$, leading to better estimation results for the zero means, and only slightly higher MSE for the nonzero means.

Thirdly, the lower MSE of the simple estimator for large values of A in case $p_n = 20$ is mostly due to a small improvement in estimating the zero means, compared to the truncated Cauchy and the MMLE. As so many of the parameters are zero, this leads to lower overall MSE. However, close to the threshold, the absolute differences between these methods on the nonzero means can be quite large.

Thus, from an estimation point of view, empirical Bayes with the MMLE or hierarchical Bayes with a truncated Cauchy seem to deliver the best results, only to be outperformed by hierarchical Bayes with a non-truncated Cauchy in a non-sparse setting with all zero means very close to the universal threshold.

3.5.2 Coverage of credible sets

We study the coverage and length of the marginal credible sets resulting from the same four methods applied in the simulation above: empirical Bayes with the simple estimator and the MMLE, and hierarchical Bayes with a Cauchy prior on τ, or a Cauchy prior truncated to $[1/n, 1]$. In addition, we study intervals of the form $\hat{\theta}_i(y_i, \widehat{\tau}_M) \pm 1.96\sqrt{\text{var}(\theta_i \mid y_i, \widehat{\tau}_M)}$, based on a normal approximation to the posterior, where $\hat{\theta}_i(y_i, \widehat{\tau}_M)$ is the posterior mean and $\text{var}(\theta_i \mid y_i, \widehat{\tau}_M)$ refers to the posterior variance, both with the MMLE plugged in. We include the approximation because it offers a computational advantage over the other methods, as no MCMC is required.

We again consider a mean vector of length $n = 400$, with $p_n \in \{20, 200\}$. We draw the nonzero means from a $\mathcal{N}(A, 1)$-distribution, with $A = c\sqrt{2 \log n}$ for $c \in \{1/2, 1, 2\}$, corresponding to most nonzero means being below the universal threshold, close to the universal threshold, or well past the universal threshold, respectively. In each of the $N = 500$ iterations, we created the 95% marginal credible sets for the hierarchical and empirical Bayes methods by taking the 2.5%- and 97.5%-quantiles of the MCMC samples as the endpoints. We did not include a blow-up factor.

Figure 3.4 gives the coverage results averaged over the 500 iterations, for all parameters, and separately for the p_n nonzero means and the $(n - p_n)$ zero means. The average lengths of the credible sets, again for all signals and separately for the nonzero and zero means, are displayed in Figure 3.5. Figure 3.6 gives the mean value of τ - in the hierarchical Bayes settings, the posterior mean of τ was recorded for each iteration. No value is given for the normal approximation, as it uses the MMLE as a plug-in value for τ.

We remark on some aspects of the results. First, we see that the zero means are nearly perfectly covered by all methods in all settings, and the main differences lie in the nonzero means. Secondly, coverage of the nonzero means improves as their values increase. Thirdly, the lengths of the credible intervals adapt to the signal size. They are smaller for the zero means than for the nonzero means, and smaller for the nonzero means corresponding to $A = (1/2)\sqrt{2 \log n}$ than for the nonzero means corresponding to

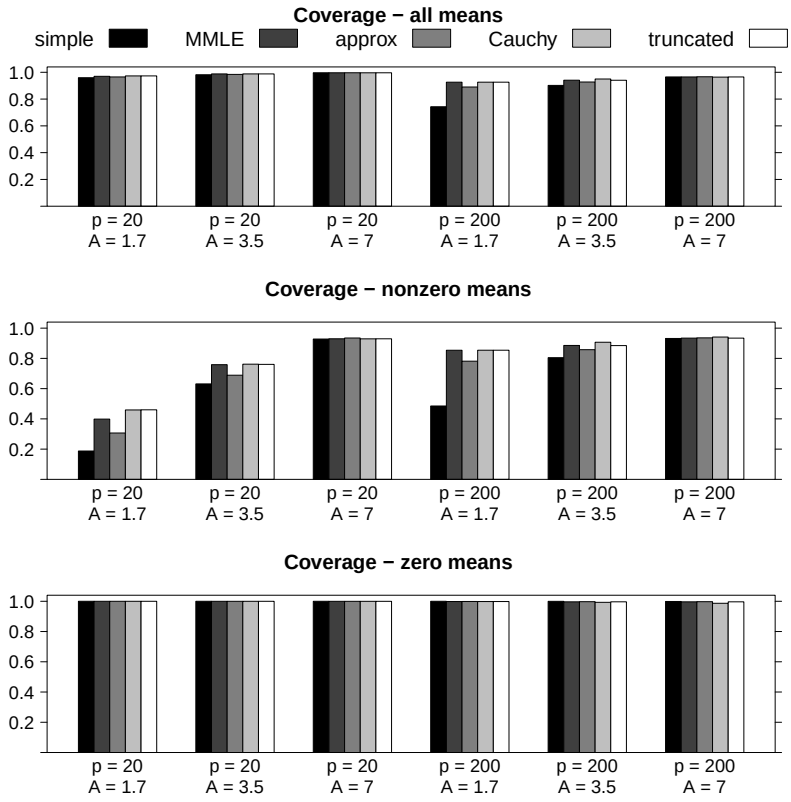

Figure 3.4: Average coverage of all parameters (top), the nonzero means (middle) and the zero means (bottom) for the five methods, from left to right: empirical Bayes with simple estimator ($c_1 = 2, c_2 = 1$) and MMLE, normal approximation, hierarchical Bayes with Cauchy prior on τ and with Cauchy prior truncated to $[1/n, 1]$. The p_n nonzero means were drawn from a $\mathcal{N}(A, 1)$ distribution. Results are based on averaging over 500 iterations.

$A = \sqrt{2 \log n}$ and $A = 2\sqrt{2 \log n}$, while there is not much difference between the interval lengths in those latter two settings, suggesting that the interval length does not increase indefinitely with the size of the nonzero mean.

Furthermore, empirical Bayes with the simple estimator achieves the lowest overall coverage, and especially bad coverage of the nonzero means. This appears to be due to smaller interval lengths caused by lower estimates of τ compared to the other methods. The normal approximation leads to better coverage than the simple estimator, and has the highest coverage of the nonzero means, even though the corresponding intervals are slightly shorter than those of empirical Bayes with the MMLE and the hierarchical Bayes approaches. However, its coverage of nonzero means is worse than that of those

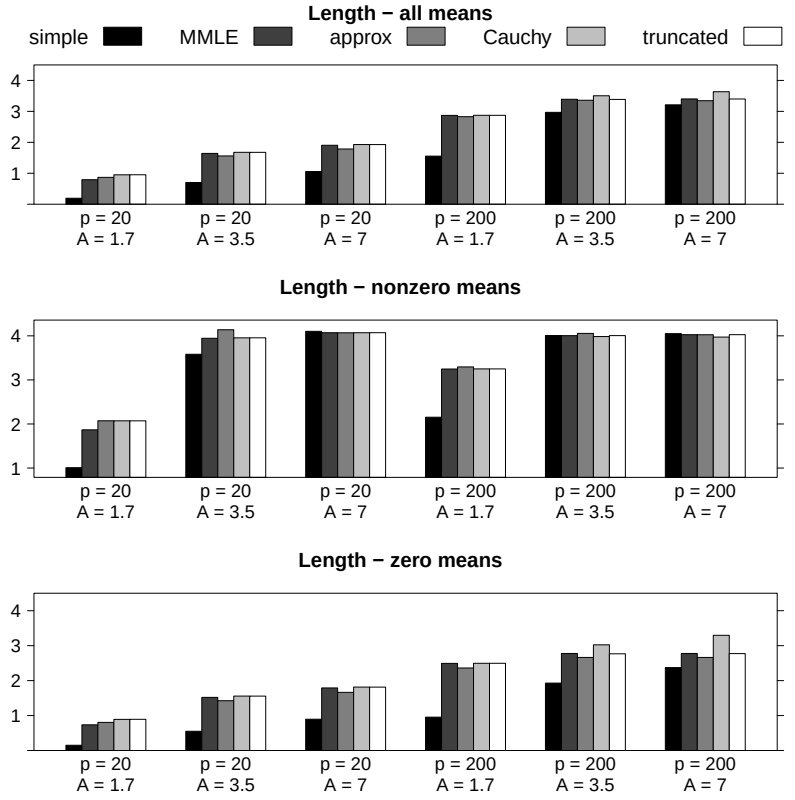

Figure 3.5: Average length of the credible sets of all parameters (top), the nonzero means (middle) and the zero means (bottom) for the five methods, from left to right: empirical Bayes with simple estimator ($c_1 = 2, c_2 = 1$) and MMLE, normal approximation, hierarchical Bayes with Cauchy prior on τ and with Cauchy prior truncated to $[1/n, 1]$. The p_n nonzero means were drawn from a $\mathcal{N}(A, 1)$ distribution. Results are based on averaging over 500 iterations.

three methods, while the corresponding intervals are longer, except in the case where A is largest. The normal approximation appears to be reasonable for very large signals only.

The hierarchical Bayes approach with a non-truncated Cauchy on τ leads to the highest overall coverage and coverage of the nonzero means, albeit by a small margin. The price is slightly larger intervals compared to the other methods, mostly for the zero means. These larger intervals are most likely due to the larger values of τ that are employed, this being the only approach that allows for estimates of τ larger than one, and it avails itself of the opportunity in the non-sparse setting. Finally, we again observe that the results for empirical Bayes with the MMLE and hierarchical Bayes with a truncated Cauchy lead to highly similar results. Their coverage is comparable to that of hierarchical Bayes with a

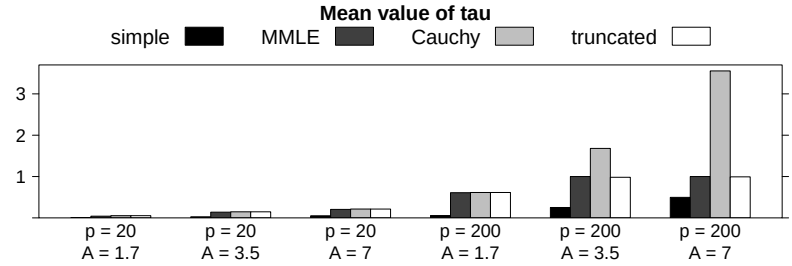

Figure 3.6: Average value of τ for four methods, from left to right: empirical Bayes with simple estimator ($c_1 = 2, c_2 = 1$) and MMLE, hierarchical Bayes with Cauchy prior on τ and with Cauchy prior truncated to $[1/n, 1]$. For the hierarchical Bayes approaches, the posterior mean of τ was recorded for each iteration. The p_n nonzero means were drawn from a $\mathcal{N}(A, 1)$ distribution. Results are based on averaging over 500 iterations. .

non-truncated Cauchy in all settings except when $p_n = 200$ and A is at least at the threshold, in which case the non-truncated Cauchy has slightly better coverage. Their intervals are shorter on average, because τ is not allowed to be larger than one.

In conclusion, empirical Bayes with the simple estimator should not be used for uncertainty quantification. The normal approximation is faster to compute than the marginal credible sets, but leads to worse coverage of the nonzero compared to the empirical Bayes with the MMLE and the hierarchical Bayes approaches, unless the nonzero means are very large. The results of those latter three methods are very similar to each other. All these results can be understood in terms of the behaviour of the estimate of τ: larger values lead to larger intervals and better coverage, which may lead to worse estimates however (as seen in the previous section). Empirical Bayes with the MMLE, or hierarchical Bayes with a truncated Cauchy, appear to be the best choices when considering both estimation and coverage. Those two approaches yield highly similar results and the choice for one over the other may be based on other considerations such as computational ones.

3.6 Proofs

3.6.1 Proofs for the main results about the MMLE

Proof of Theorem 3.1

Proof. By its definition the MMLE maximizes the logarithm of the marginal likelihood function, which is given by

$$M_\tau(Y^n) = \sum_{i=1}^{n} \log\Big(\int_{-\infty}^{\infty} \varphi(y_i - \theta) g_\tau(\theta) d\theta \Big). \tag{3.18}$$

We split the sum in the indices $I_0 := \{i : \theta_{0,i} = 0\}$ and $I_1 := \{i : \theta_{0,i} \neq 0\}$. By Lemma 3.22, with m_τ given by (3.37),

$$\frac{d}{d\tau} M_\tau(Y^n) = \frac{1}{\tau} \sum_{i \in I_0} m_\tau(Y_i) + \frac{1}{\tau} \sum_{i \in I_1} m_\tau(Y_i).$$

By Proposition 3.23 the expectations of the terms in the first sum are strictly negative and bounded away from zero for $\tau \geq \varepsilon$, and any given $\varepsilon > 0$. By Lemma 3.27 the sum behaves likes its expectation, uniformly in τ. By Lemma 3.28 (i) the function m_τ is uniformly bounded by a constant C_u. It follows that for every $\varepsilon > 0$ there exists a constant $C_\varepsilon > 0$ such that, for all $\tau \geq \varepsilon$, and with $p_n = \#(\theta_{0,i} \neq 0)$, the preceding display is bounded above by

$$-\frac{n - p_n}{\tau} C_\varepsilon (1 + o_P(1)) + \frac{p_n}{\tau} C_u.$$

This is negative with probability tending to one as soon as $(n - p_n)/p_n > C_u/C_e$, and in that case the maximum $\widehat{\tau}_M$ of $M_\tau(Y^n)$ is taken on $[1/n, \varepsilon]$. Since this is true for any $\varepsilon > 0$, we conclude that $\widehat{\tau}_M$ tends to zero in probability.

We can now apply Proposition 3.23 and Lemma 3.24 to obtain the more precise bound on the derivative when $\tau \to 0$ given by

$$\frac{d}{d\tau} M_\tau(Y^n) \leq -\frac{(n - p_n)(2/\pi)^{3/2}}{\zeta_\tau} (1 + o_P(1)) + \frac{p_n}{\tau} C_u. \tag{3.19}$$

This is negative for $\tau/\zeta_\tau \gtrsim p_n/(n - p_n)$, and then $\widehat{\tau}_M$ is situated on the left side of the solution to this equation, or $\widehat{\tau}_M/\zeta_{\widehat{\tau}_M} \lesssim p_n/(n - p_n)$, which implies, that $\widehat{\tau}_M \lesssim \tau_n$, given the assumption that $p_n = o(n)$. □

Proof of Lemma 3.12

Proof. Given θ_0 that satisfies the excessive-bias restriction, let $\tilde{\zeta} = A\sqrt{2\log(n/q)}$ and $\tilde{p} = \#\big(i : |\theta_{0,i}| \geq \tilde{\zeta}\big)$, for q as in (3.14). Then $q/C_s \leq \tilde{p} \leq p = \#\big(i : \theta_{0,i} \neq 0\big) \leq p_n$, which is $o(n)$ by assumption, so that $\tilde{\zeta} \to \infty$, uniformly in θ_0.

Take any $\delta_n \downarrow 0$ and $A_1 \in (A^{-1}, 1)$ and for given τ split the set of indices $1, \ldots, n$ into $I_2 := \{i : |Y_i| \geq A_1\tilde{\zeta}\}$, $I_0 = \{i \notin I_2 : |\theta_{0,i}| \leq \delta_n \zeta_\tau^{-2}\}$, and $I_1 = I_2^c \cap I_0^c$ the remaining indices. Since $|Y_i| \geq |\theta_{0,i}| - |\varepsilon_i|$, we have that $i \in I_2$ as soon as $|\theta_{0,i}| \geq \tilde{\zeta}$ and $|\varepsilon_i| < (1 - A_1)\tilde{\zeta}$. By definition there exist $\tilde{p}$ coordinates with $|\theta_{0,i}| \geq \tilde{\zeta}$, and the number of the corresponding variables $|\varepsilon_i|$ that fall below $(1 - A_1)\tilde{\zeta}$ is a binomial variable on $\tilde{p}$ trials and success probability tending to one, as $(1 - A_1)\tilde{\zeta} \to \infty$. By Chebyshev's inequality it follows that with probability tending to one the cardinality of I_2 is at least $\tilde{p}/2$ (easily). By the excessive-bias restriction

$$\delta_n^2 \zeta_\tau^{-4} \#\big(i : \delta_n \zeta_\tau^{-2} < |\theta_{0,i}| < \tilde{\zeta}\big) \leq \sum_{i : |\theta_{0,i}| < \tilde{\zeta}} \theta_{0,i}^2 \lesssim q \log(n/q) \leq C_s \tilde{p} \log(ne/(C_s\tilde{p})).$$

This shows that the number of elements of I_1 with $|\theta_{0,i}| < \tilde{\zeta}$ is bounded above by a multiple of $\delta_n^{-2}\zeta_\tau^4 \tilde{p} \log(ne/(C_s\tilde{p}))$. The number of $\theta_{0,i}$ with $|\theta_{0,i}| \geq \tilde{\zeta}$ is $\tilde{p}$ by definition, which is

smaller than the preceding number if δ_n tends to zero sufficiently slowly and ζ_τ is bounded away from 0. In that case the cardinality of I_1 is bounded above by $\delta_n^{-2}\zeta_\tau^4\tilde{p}\log(ne/(C_s\tilde{p}))$. Since the indices of all zero coordinates are contained in I_0, the cardinality of I_1 is also trivially bounded from above by p.

By Lemma 3.22 the derivative of the log-likelihood can be written in the form

$$\begin{aligned}\frac{d}{d\tau}M_\tau(Y^n) &= \frac{1}{\tau}\sum_{i\in I_0} m_\tau(Y_i) + \frac{1}{\tau}\sum_{i\in I_1} m_\tau(Y_i) + \frac{1}{\tau}\sum_{i\in I_2} m_\tau(Y_i) \qquad (3.20)\\ &\geq -\frac{C_e}{\zeta_\tau}n - |I_1| + |I_2|C\Big(\frac{1}{\tau}\wedge\frac{e^{A_1^2\tilde{\zeta}^2/2}}{A_1^2\tilde{\zeta}^2}\Big),\end{aligned}$$

with probability tending to 1, uniformly in $\tau \in [1/n, \eta_n]$ and any $\eta_n \downarrow 0$, for constants $C_e, C > 0$. This follows by applying Proposition 3.23 together with Lemma 3.24 to the first sum, Lemma 3.28(ii) and the monotonicity of $y \mapsto m_\tau(y)$ to the second, and Lemma 3.28(vi) to the third sum. The right side is certainly nonnegative for τ such that the third term dominates twice the absolute values of both the first and second terms. Since $|I_2| \geq \tilde{p}/2$ and $\tilde{p} \gtrsim q = ne^{-\tilde{\zeta}^2A^{-2}/2}$, it follows that the right side is nonnegative if

$$\frac{n}{\zeta_\tau} \lesssim \frac{\tilde{p}}{\tau}, \qquad \frac{n}{\zeta_\tau} \lesssim \frac{ne^{\tilde{\zeta}^2(A_1^2-A^{-2})/2}}{\tilde{\zeta}^2}, \qquad |I_1| \lesssim \frac{\tilde{p}}{\tau}, \qquad |I_1| \lesssim \frac{ne^{\tilde{\zeta}^2(A_1^2-A^{-2})/2}}{\tilde{\zeta}^2},$$

where the multiplicative constants must be sufficiently small. The first inequality is satisfied for $\tau \lesssim \tau_n(\tilde{p})$; the second is trivial since $A_1 > A^{-1}$ and $\tilde{\zeta} \to \infty$, and $\zeta_\tau^{-1} \to 0$; the third can be reduced to $\tau\zeta_\tau^4 \lesssim \delta_n^2/\log(ne/(C_s\tilde{p}))$, which is (easily) verified if $\tau \lesssim \tau_n(\tilde{p})$ and δ_n tends to zero sufficiently slowly; the fourth is trivial since $|I_1| \leq p \ll n$ and $A_1 > A^{-1}$ and $\tilde{\zeta} \to \infty$. It follows that $\tau \mapsto M_\tau(Y^n)$ is increasing for $\tau \lesssim \tau_n(\tilde{p})$ and hence $\widehat{\tau}_M \gtrsim \tau_n(\tilde{p})$.

For the proof of the upper bound we use the same decomposition (3.20), but redefine the sets I_k slightly, to $I_0 = \{i : |\theta_{0,i}| \leq \delta_n/\zeta_\tau^2\}$, $I_1 = \{\delta_n/\zeta_\tau^2 \leq |\theta_{0,i}| \leq \zeta_\tau/4\}$ and $I_2 = I_0^c \cap I_1^c$. Reasoning as before, using the excessive-bias restriction, we see that the cardinalities of the sets I_1 and I_2 are bounded by multiples of $\delta_n^{-2}\zeta_\tau^4\tilde{p}\log(ne/(C_s\tilde{p}))$ and $\zeta_\tau^{-2}\tilde{p}\log(ne/(C_s\tilde{p})) + \tilde{p}$, respectively. By the decomposition (3.20) we obtain,

$$\frac{d}{d\tau}M_\tau(Y^n) \lesssim -\frac{C_e}{\zeta_\tau}(n-p) + \frac{|I_1|\tau^{1/16}}{\tau\zeta_\tau} + o\Big(\frac{|I_1|\tau^{1/32}}{\tau\zeta_\tau}\Big) + \frac{1}{\tau}|I_2|C_u, \qquad (3.21)$$

with probability tending to 1, uniformly in $\tau \in [1/n, \eta_n]$ and any $\eta_n \downarrow 0$. Here the upper bounds on the sums over the coordinates in I_0 and I_1 follow with the help of the first and second parts of Proposition 3.23 and Lemma 3.24, and the bound on the sum over the coordinates in I_2 follows from Lemma 3.28(i). The right side is certainly negative for τ such that $2\tau^{-1}|I_2|C_u \leq C_e(n-p)/\zeta_\tau$ and $|I_1|\tau^{1/32}/\zeta_\tau \leq C_u|I_2|$. The first reduces to $\tau\zeta_\tau \gtrsim (\tilde{p}/n)\log(ne/(C_s\tilde{p}))$ and $\tau/\zeta_\tau \gtrsim \tilde{p}/n$ and hence is true for $\tau \gtrsim \tau_n(\tilde{p})$; the second reduces to $\tau^{1/32}\zeta_\tau^5 \lesssim \delta_n^2$ and is true as well provided $\delta_n \downarrow 0$ slowly. Since we may assume that $\widehat{\tau}_M \in [1/n, \eta_n]$ for some $\eta_n \downarrow 0$ by Theorem 3.1, it follows in that case that $\widehat{\tau}_M \lesssim \tau_n(\tilde{p})$. □

3.6.2 Proofs of the contraction results

Lemma 3.16. For $A > 1$ and every $y \in \mathbb{R}$,

(i) $|\mathbb{E}(\theta_i \mid Y_i = y, \tau) - y| \le 2\zeta_\tau^{-1}$, for $|y| \ge A\zeta_\tau$, as $\tau \to 0$.

(ii) $|\mathbb{E}(\theta_i \mid Y_i = y, \tau)| \le |y|$.

(iii) $|\mathbb{E}(\theta_i \mid Y_i = y, \tau)| \le \tau |y| e^{y^2/2}$, as $\tau \to 0$.

(iv) $|\operatorname{var}(\theta_i \mid Y_i = y, \tau) - 1| \le \zeta_\tau^{-2}$, for $|y| \ge A\zeta_\tau$, as $\tau \to 0$.

(v) $\operatorname{var}(\theta_i \mid Y_i = y, \tau) \le 1 + y^2$,

(vi) $\operatorname{var}(\theta_i \mid Y_i = y, \tau) \lesssim \tau e^{y^2/2}(y^{-2} \wedge 1)$, as $\tau \to 0$.

(vii) $|\mathbb{E}(\theta_i \mid Y_i = y, \tau) - y| \lesssim (\log |y|)/|y|$, uniformly in $\tau \ge \tau_0 > 0$ and $|y| \to \infty$.

Proof. Inequalities (iii) and (v) come from Lemma 1.8 and Lemma 1.10 in Chapter 1, while (ii), (iv) and (vi) are implicit in the proofs of Theorems 1.1 and 1.2 (twice) in Chapter 1, and (i) with the bound ζ_τ instead of ζ_τ^{-1} is (1.17) there. Alternatively, the posterior mean and variance in these assertions are given in (3.25) and (3.26). Then (ii) and (iv) are immediate from the fact that $0 \le I_{3/2} \le I_{1/2} \le I_{-1/2}$, while (iii) and (vi) follow by bounding $I_{-1/2}$ below by a multiple of $1/\tau$ and $I_{3/2} \le I_{1/2}$ above by $(1 \wedge y^{-2})e^{y^2/2}$, using Lemmas 3.30 and 3.31. Assertions (i) and (iv) follow from expanding $I_{-1/2}$ and $I_{1/2}$ and $I_{3/2}$, again using Lemmas 3.30 and 3.31. Finally (vii) follows from Lemma 3.32. □

3.6.3 Proof of Theorem 3.2

Proof. Set $r_n = \sqrt{p_n \log n}$ and $\tau_n = \tau_n(p_n)$. By Condition 4 and the triangle inequality,

$$\begin{aligned}
&\mathbb{E}_{\theta_0} \Pi_{\widehat{\tau}_n}\Big(\theta : \|\theta_0 - \theta\|_2 \ge M_n r_n \mid Y^n\Big) \\
&\quad \le \mathbb{E}_{\theta_0} \mathbf{1}_{\widehat{\tau}_n \in [1/n, C\tau_n]} \Pi_{\widehat{\tau}_n}\Big(\theta : \|\theta_0 - \hat\theta(\widehat{\tau}_n)\|_2 + \|\theta - \hat\theta(\widehat{\tau}_n)\|_2 \ge M_n r_n \mid Y^n\Big) + o(1) \\
&\quad \le \mathbb{E}_{\theta_0} \sup_{\tau \in [1/n, C\tau_n]} \Pi_\tau\Big(\theta : \|\theta_0 - \hat\theta(\tau)\|_2 + \|\theta - \hat\theta(\tau)\|_2 \ge M_n r_n \mid Y^n\Big) + o(1).
\end{aligned}$$

Hence, in view of Chebyshev's inequality, it is sufficient to show that, with $\operatorname{var}(\theta \mid Y^n, \tau) = \mathbb{E}\Big(\|\theta - \hat\theta(\tau)\|^2 \mid Y^n, \tau\Big)$,

$$P_{\theta_0}\Big(\sup_{\tau \in [1/n, C\tau_n]} \|\theta_0 - \hat\theta(\tau)\|_2 \ge (M_n/2) r_n\Big) = o(1), \tag{3.22}$$

$$P_{\theta_0}\Big(\sup_{\tau \in [1/n, C\tau_n]} \operatorname{var}(\theta \mid Y^n, \tau) \ge M_n r_n^2\Big) = o(1). \tag{3.23}$$

To prove (3.22) we first use Lemma 3.16(i)+(ii) to see that $|\hat\theta_i(\tau)| \lesssim \zeta_\tau$ and next the triangle inequality to see that $|\hat\theta_i(\tau) - \theta_{0,i}| \lesssim \zeta_\tau + |Y_i - \theta_{0,i}|$, as $\tau \to 0$. This shows that

$$\mathbb{E}_{\theta_{0,i}} \sup_{\tau \in [1/n, \tau_n]} (\theta_{0,i} - \hat\theta_i(\tau))^2 \lesssim \sup_{\tau \ge 1/n} \zeta_\tau^2 + \operatorname{var}_{\theta_{0,i}} Y_i \lesssim \log n. \tag{3.24}$$

Second we use Lemma 3.16 (iii) and (ii) to see that $|\hat{\theta}_i(\tau)|$ is bounded above by $\tau|Y_i|e^{Y_i^2/2}$ if $|Y_i| \le \zeta_{\tau_n}$ and bounded above by $|Y_i|$ otherwise, so that

$$\mathbb{E}_0 \sup_{\tau\in[1/n,C\tau_n]} |\hat{\theta}_i(\tau)|^2 \lesssim \int_0^{\zeta_{\tau_n}} (C\tau_n)^2 y^2 e^{y^2}\varphi(y)\,dy + \int_{\zeta_{\tau_n}}^{\infty} y^2\varphi(y)\,dy \lesssim \tau_n\zeta_{\tau_n}.$$

Applying the upper bound (3.24) for the p_n non-zero coordinates $\theta_{0,i}$, and the upper bound in the last display for the zero parameters, we find that

$$\mathbb{E}_{\theta_0} \sup_{\tau\in[1/n,C\tau_n]} \|\theta_0 - \hat{\theta}(\tau)\|_2^2 \lesssim p_n \log n + (n-p_n)\tau_n\zeta_{\tau_n} \lesssim p_n \log n.$$

Next an application of Markov's inequality leads to (3.22).

The proof of (3.23) is similar. For the nonzero $\theta_{0,i}$ we use the fact that $\operatorname{var}(\theta_i \mid Y_i, \tau) \le 1 + \zeta_\tau^2 \lesssim \log n$, by Lemma 3.16 (iv) and (v), while for the zero $\theta_{0,i}$ we use that $\operatorname{var}(\theta_i \mid Y_i, \tau)$ is bounded above by $\tau e^{Y_i^2/2}$ for $|Y_i| \le \zeta_{\tau_n}$ and bounded above by $1 + Y_i^2$ otherwise, by Lemma 3.16 (vi) and (v). For the two cases of parameter values this gives bounds for $\mathbb{E}_{\theta_{0,i}} \sup_{\tau\in[1/n,C\tau_n]} \operatorname{var}(\theta_i \mid Y_i, \tau)$ of the same form as the bounds for the square bias, resulting in the overall bound $p_n \log n + (n-p_n)\tau_n\zeta_{\tau_n} \lesssim p_n \log n$ for the sum of these variances. An application of Markov's inequality gives (3.23). □

Proof of Lemma 3.6

Proof. The number t_n defined in Condition 7 is the (approximate) solution to the equation p_nC_u/τ
$= C_e(n-p)/(2\zeta_\tau)$, for $C_e = (\pi/2)^{3/2}$. By the decomposition (3.19), with P_{θ_0}-probability tending to one,

$$\frac{\partial}{\partial\tau} M_\tau(Y^n) < \begin{cases} p_nC_u/(t_n/2), & \text{if } t_n/2 \le \tau \le t_n, \\ 0 & \text{if } \tau > t_n, \\ -p_nC_u/(2t_n), & \text{if } \tau \ge 2t_n. \end{cases}$$

Therefore, for $M_\tau(Y^n)$ defined in (3.18), $\tau_{\min} = \operatorname{argmin}_{\tau\in[t_n/2,t_n]} M_\tau(Y^n)$, and $\tau \ge 2t_n$,

$$\begin{aligned} M_\tau(Y^n) - M_{\tau_{\min}}(Y^n) &= \Big[\int_{\tau_{\min}}^{t_n} + \int_{t_n}^{2t_n} + \int_{2t_n}^{\tau}\Big] \frac{\partial}{\partial s} M_s(Y^n)\,ds \\ &\le (t_n/2)p_nC_u/(t_n/2) + 0 - (\tau - 2t_n)p_nC_u/(2t_n) \\ &= -(\tau - 4t_n)p_nC_u/(2t_n) \le -\tau p_nC_u/(10t_n), \end{aligned}$$

for $\tau \ge 5t_n$. Since $\pi(\tau \mid Y^n) \propto \pi(\tau)e^{M_\tau(Y^n)}$ by Bayes's formula, with P_{θ_0}-probability tending to one, for $c_n \ge 5$

$$\Pi(\tau \ge c_nt_n \mid Y^n) \le \frac{\int_{\tau\ge c_nt_n} e^{M_{\tau_{\min}}(Y^n) - \tau p_nC_u/(10t_n)}\pi(\tau)\,d\tau}{\int_{\tau\in[t_n/2,t_n]} e^{M_{\tau_{\min}}(Y^n)}\pi(\tau)\,d\tau} \lesssim \frac{e^{-c_np_nC_u/10}}{\int_{\tau\in[t_n/2,t_n]}\pi(\tau)\,d\tau}.$$

Under Condition 6 this tends to zero if $c_n \ge 5$. Under the weaker Condition 7 this is certainly true for $c_n \ge \log n$. □

3.6.4 Proofs for the coverage of the credible sets

Proof of Lemma 3.8

Proof. The square radius $\hat{r}^2(\alpha,\tau)$ is defined as the upper α-quantile of the variable $W = \|\theta - \hat{\theta}(\tau)\|_2^2$ relative to its posterior distribution given (Y^n,τ), where $\hat{\theta}(\tau) = \mathbb{E}(\theta \mid Y^n,\tau)$. By Chebyshev's inequality the variable W falls below $\mathbb{E}(W \mid Y^n,\tau) - c\,\mathrm{sd}(W \mid Y^n,\tau)$ with conditional probability given (Y^n,τ) smaller than $1/c^2$ for any given $c > 0$. This implies that $\hat{r}^2(\alpha,\tau) \geq \mathbb{E}(W \mid Y^n,\tau) - c\,\mathrm{sd}(W \mid Y^n,\tau)$ for $c > 0$ such that $1/c^2 \leq 1-\alpha$. Thus it suffices to show that $\mathbb{E}(W \mid Y^n,\tau) \geq 0.501 n\tau\zeta_\tau$ and $\mathrm{sd}(W \mid Y^n,\tau) \ll n\tau\zeta_\tau$, with probability tending to 1. Here the conditional expectations $\mathbb{E}(W \mid Y^n,\tau)$ and $\mathrm{sd}(W \mid Y^n,\tau)$ refer to the posterior distribution of θ given (Y^n,τ) (where W is a function of θ), which are functions of Y^n that will be considered under the law of Y^n following the true parameter. The variable $W = \sum_{i=1}^n \big(\theta_i - \hat{\theta}_i(\tau)\big)^2$ is lower bounded by the sum of squares W_0 of the variables $\theta_i - \hat{\theta}_i(\tau)$ corresponding to the indices with $\theta_{0,i} = 0$, which are $(n - p_n) \sim n$ of the coordinates. The upper α-quantile of W is bigger than the upper α-quantile of W_0, and hence it suffices to derive a lower bound for the latter. For simplicity of notation we assume that all n parameters $\theta_{0,i}$ are zero and write W for W_0.

Because given τ the coordinates are independent under the posterior distribution,

$$\mathbb{E}(W \mid Y^n,\tau) = \sum_{i=1}^n \mathbb{E}\Big[\big(\theta_i - \hat{\theta}_i(\tau)\big)^2 \mid Y_i,\tau\Big] = \sum_{i=1}^n \mathrm{var}(\theta_i \mid Y_i,\tau),$$
$$\mathrm{var}(W \mid Y^n,\tau) = \sum_{i=1}^n \mathrm{var}\Big[\big(\theta_i - \hat{\theta}_i(\tau)\big)^2 \mid Y_i,\tau\Big] \leq \sum_{i=1}^n \mathbb{E}\Big[\big(\theta_i - \hat{\theta}_i(\tau)\big)^4 \mid Y_i,\tau\Big].$$

Because the variables Y_i are i.i.d. under the true distribution, Lemma 3.17 below gives that

$$\mathbb{E}_0\mathbb{E}(W \mid Y^n,\tau) \sim (2/\pi)^{3/2} n\tau\zeta_\tau,$$
$$\mathrm{var}_0\,\mathbb{E}(W \mid Y^n,\tau) \lesssim n\tau\zeta_\tau,$$
$$\mathbb{E}_0\,\mathrm{var}(W \mid Y^n,\tau) \lesssim n\tau\zeta_\tau^3.$$

From the first two assertions and another application of Chebyshev's inequality, now with respect to the true law of Y^n, it follows that for any $c_n \to \infty$ the probability of the event $\mathbb{E}(W \mid Y^n,\tau) \leq (2/\pi)^{3/2} n\tau\zeta_\tau - c_n\sqrt{n\tau\zeta_\tau}$ tends to zero. Since $\sqrt{n\tau\zeta_\tau} \ll n\tau\zeta_\tau$ (easily) under the assumption that $n\tau/\zeta_\tau \to \infty$ and $(2/\pi)^{3/2} \approx 0.507$, it follows that $\mathbb{E}(W \mid Y^n,\tau)$ is lower bounded by $0.5 n\tau\zeta_\tau$ with probability tending to one. By Markov's inequality the probability of the event $\mathrm{sd}(W \mid Y^n,\tau) \geq c_n n\tau\zeta_\tau$ is bounded above by $(c_n n\tau\zeta_\tau)^{-2}\mathbb{E}_0\,\mathrm{var}(W \mid Y^n,\tau)$, which is further bounded above by $(c_n n\tau\zeta_\tau)^{-2} n\tau\zeta_\tau^3$, by the third assertion in the display. This tends to zero for some $c_n \to 0$, again by the assumption that $n\tau/\zeta_\tau \to \infty$ (tightly this time). □

For the proof of Lemma 3.8, we have employed the lemma below, which is based on the following observations. The posterior density of θ_i given $(Y_i = y,\tau)$ is (for fixed τ) an exponential family with density

$$\theta \mapsto \frac{\varphi(y-\theta)g_\tau(\theta)}{\psi_\tau(y)} = c_\tau(y)e^{\theta y}g_\tau(\theta)e^{-\theta^2/2},$$

where g_τ is the prior density of θ given in (3.3), and ψ_τ is the Bayesian marginal density of Y_i, given in (3.36), and the norming constant is given by

$$c_\tau(y) = \frac{\varphi(y)}{\psi_\tau(y)} = \frac{\pi}{\tau I_{-1/2}(y)},$$

for the function $I_{-1/2}(y)$ defined in (3.35). The cumulant moment generating function $z \mapsto \log \mathbb{E}(e^{z\theta_i} \mid Y_i = y, \tau)$ of the family is given by $z \mapsto \log\bigl(c_\tau(y)/c_\tau(y+z)\bigr)$, which is $z \mapsto \log I_{-1/2}(y+z)$ plus an additive constant independent of z. We conclude that the first, second and fourth cumulants are given by

$$\begin{aligned} \hat\theta_i(\tau) = \mathbb{E}(\theta_i \mid Y_i = y, \tau) &= \frac{d}{dy} \log I_{-1/2}(y), \\ \operatorname{var}(\theta_i \mid Y_i = y, \tau) &= \frac{d^2}{dy^2} \log I_{-1/2}(y), \\ \mathbb{E}\Bigl[\bigl(\theta_i - \hat\theta_i(\tau)\bigr)^4 \mid Y_i = y, \tau\Bigr] - 3\operatorname{var}(\theta_i \mid Y_i = y, \tau)^2 &= \frac{d^4}{dy^4} \log I_{-1/2}(y). \end{aligned} \tag{3.25}$$

The derivatives at the right side can be computed by repeatedly using the product and sum rule together with the identity $I_k'(y) = yI_{k+1}(y)$, for I_k as in (3.35).

Lemma 3.17. For $\mathbb{E}_0$ referring to the distribution of $Y_i \sim N(0,1)$, as $\tau \to 0$,

$$\begin{aligned} \frac{4C^{-1}\tau\zeta_\tau}{\pi\sqrt{2\pi}} \lesssim \mathbb{E}_0 \inf_{t\in[C^{-1}\tau, C\tau]} \operatorname{var}(\theta_i \mid Y_i, t) \lesssim \mathbb{E}_0 \sup_{t\in[C^{-1}\tau, C\tau]} \operatorname{var}(\theta_i \mid Y_i, t) \lesssim \frac{4C\tau\zeta_\tau}{\pi\sqrt{2\pi}}, \\ \mathbb{E}_0 \sup_{t\in[C^{-1}\tau, C\tau]} \operatorname{var}(\theta_i \mid Y_i, t)^2 \lesssim \tau\zeta_\tau, \\ \mathbb{E}_0 \sup_{t\in[C^{-1}\tau, C\tau]} \mathbb{E}\Bigl[\bigl(\theta_i - \hat\theta_i(t)\bigr)^4 \mid Y_i, t\Bigr] \lesssim \tau\zeta_\tau^3. \end{aligned}$$

Proof. The first assertion is already contained in Chapter 1, but we give a new proof, which also prepares for the proofs of the other assertions.

Since $(\log h)'' = h''/h - (h'/h)^2$, for any function h, and $I_{-1/2}'(y) = yI_{1/2}(y)$ and $I_{-1/2}''(y) = y^2 I_{3/2}(y) + I_{1/2}(y)$, we have by the formulas preceding the lemma,

$$\operatorname{var}(\theta_i \mid Y_i = y, \tau) = y^2\Bigl[\frac{I_{3/2}}{I_{-1/2}} - \Bigl(\frac{I_{1/2}}{I_{-1/2}}\Bigr)^2\Bigr](y) + \frac{I_{1/2}}{I_{-1/2}}(y). \tag{3.26}$$

By Lemmas 3.30 and 3.31 the right side is equivalent, uniformly in y, to

$$\begin{aligned} y^2\Bigl[\frac{H_{3/2}(y)}{\pi/\tau + H_{-1/2}(y)}\bigl(1 + O(\sqrt\tau)\bigr) - \frac{H_{1/2}^2(y)}{\bigl(\pi/\tau + H_{-1/2}(y)\bigr)^2}\bigl(1 + O(\sqrt\tau)\bigr)\Bigr] \\ + \frac{H_{1/2}(y)}{\pi/\tau + H_{-1/2}(y)}\bigl(1 + O(\sqrt\tau)\bigr), \end{aligned}$$

where $H_k(y) = (y^2/2)^{-k}\int_c^{y^2/2} v^{k-1}e^v\,dv$, with $c = 0$ if $k > 0$ and $c = 1$ otherwise. Uniformly in $y \geq 1/\varepsilon_\tau \to \infty$, all functions H_k can be expanded as $H_k(y) = e^{y^2/2}/(y^2/2)(1+O(1/y^2))$, by Lemma 3.29.

Let κ_τ be the solution to $e^{\kappa_\tau^2/2}/(\kappa_\tau^2/2) = 1/\tau$. For $y \ll \kappa_\tau$ the factor π/τ dominates the factor $H_{-1/2}(y)$ and the preceding display can be approximated by

$$\frac{\tau}{\pi}y^2H_{3/2}(y) - \frac{\tau^2}{\pi^2}y^2H_{1/2}^2(y) + \frac{\tau}{\pi}H_{1/2}(y). \tag{3.27}$$

For instance, we can use this approximation on $[0,\zeta_\tau]$, up to a uniform $1+o(1)$-term, since $e^{-\zeta_\tau^2/2}/\zeta_\tau^2 \ll 1/\tau$. A multiple of the preceding display, with the negative term removed, is an upper bound for $\operatorname{var}(\theta_i \mid Y_i,\tau)$ for any y; we use this for $y \in [\zeta_\tau,\kappa_\tau]$. For $y \geq \kappa_\tau$ the factor $H_{-1/2}(y)$ dominates π/τ and the second to last display can be rewritten as, for $\delta_\tau(y) = (\pi/\tau)/H_{-1/2}(y)$,

$$\begin{aligned} y^2\Big[\frac{1+O(y^{-2})}{1+\delta_\tau(y)}(1+o(1)) - \frac{1+O(y^{-2})}{(1+\delta_\tau(y))^2}(1+o(1))\Big] + \frac{1+O(y^{-2})}{1+\delta_\tau(y)}(1+o(1)) \\ = \frac{y^2\delta_\tau(y)}{(1+\delta_\tau(y))^2} + r_\tau(y), \end{aligned} \tag{3.28}$$

where $r_\tau(y)$ is uniformly bounded in $y \geq \kappa_\tau$ as $\tau \to 0$.

We can choose $\varepsilon_{\tau/C} \to 0$ slow enough that

$$\mathbb{E}_0 \sup_{t\in[C^{-1}\tau,C\tau]} \operatorname{var}(\theta_i \mid Y_i,t)\mathbf{1}_{0\leq|Y_i|\leq 1/\varepsilon_t} \lesssim C\tau\int_0^{1/\varepsilon_{\tau/C}} \Big[y^2H_{3/2}(y) + H_{1/2}(y)\Big]\varphi(y)\,dy$$

is of smaller order than $\tau\zeta_\tau$. Then this part of the expectation is negligible. For $1/\varepsilon_t \leq |y| \leq \zeta_t$, we expand the functions H_k in (3.27) and find that

$$\begin{aligned} &\mathbb{E}_0 \sup_{t\in[C^{-1}\tau,C\tau]} \operatorname{var}(\theta_i \mid Y_i,t)\mathbf{1}_{1/\varepsilon_t\leq|Y_i|\leq\zeta_\tau} \\ &\quad\lesssim 2\int_0^\infty \sup_{t\in[C^{-1}\tau,C\tau]} \mathbf{1}_{1/\varepsilon_t\leq|y|\leq\zeta_\tau}\Big[(2\tau/\pi)e^{y^2/2} \\ &\qquad - (2\tau/\pi)^2e^{y^2}/y^2 + (2\tau/\pi)e^{y^2/2}/y^2\Big]\varphi(y)\,dy. \end{aligned} \tag{3.29}$$

We note that the integrand is non-negative and its derivative with respect to t is also non-negative for every $1/\varepsilon_{C\tau} \leq y \leq \zeta_{\tau/C}$ and $t \leq C\tau$, i.e.

$$(2/\pi)e^{y^2/2} - (8\tau/\pi^2)e^{y^2}/y^2 + (2/\pi)e^{y^2/2}/y^2 > 0,$$

since $e^{y^2/2} \leq C/\tau$ and $y^2 \to \infty$. Therefore, we can further bound the right hand side of (3.29) as

$$2\int_{\varepsilon_{C\tau}^{-1}}^{\zeta_{\tau/C}} \Big[(2C\tau/\pi)e^{y^2/2} - (2C\tau/\pi)^2e^{y^2}/y^2 + (2C\tau/\pi)e^{y^2/2}/y^2\Big]\varphi(y) \asymp \frac{2C\sqrt{2}}{\pi\sqrt{\pi}}\tau\zeta_\tau.$$

Similar computations also lead to

$$\mathbb{E}_0 \inf_{t\in[C^{-1}\tau,C\tau]} \operatorname{var}(\theta_i \mid Y_i,t)\mathbf{1}_{1/\varepsilon_t\le|Y_i|\le\zeta_\tau} \gtrsim \frac{2C\sqrt{2}}{\pi\sqrt{\pi}}\tau\zeta_\tau.$$

For $y \in [\zeta_\tau,\kappa_\tau]$ we again use (3.27), but as an upper bound (without the negative term), and obtain

$$\mathbb{E}_0 \sup_{t\in[C^{-1}\tau,C\tau]} \operatorname{var}(\theta_i \mid Y_i,t)\mathbf{1}_{\zeta_\tau\le|Y_i|\le\kappa_\tau} \lesssim C\tau \int_{\zeta_{C\tau}}^{\kappa_{C^{-1}\tau}} e^{y^2/2}\varphi(y)\,dy \lesssim \tau(\kappa_\tau-\zeta_\tau),$$

which is of lower order than the preceding display. By (3.28) the contribution of $y \ge \kappa_\tau$ is bounded by

$$\begin{aligned}\mathbb{E}_0 \sup_{t\in[C^{-1}\tau,C\tau]} \operatorname{var}(\theta_i \mid Y_i,t)\mathbf{1}_{\kappa_\tau\le|Y_i|} &\lesssim \int_{\kappa_{\tau/C}}^{\infty} [y^2\delta_{\tau/C}(y)+1]\varphi(y)\,dy\\ &\lesssim \int_{\kappa_{\tau/C}}^{\infty} [C\tau^{-1}y^4e^{-y^2}+e^{-y^2/2}]\,dy\\ &\lesssim \tau^{-1}\kappa_\tau^3e^{-\kappa_\tau^2}+\kappa_\tau^{-1}e^{-\kappa_\tau^2/2} = O(\tau/\kappa_\tau).\end{aligned}$$

This concludes the proof of the first assertion.

For the proof of the second assertion we follow the same approach. We simply square the integrands in the preceding bounds and obtain a negligible contribution from the interval $[0,1/\varepsilon_\tau]$, a contribution bounded by $C^2\tau^2\int_0^{\kappa_{\tau/C}} e^{y^2}\varphi(y)\,dy \lesssim \tau^2e^{\kappa_{\tau/C}^2/2}/\kappa_\tau \lesssim \tau\zeta_\tau$ from the interval $[1/\varepsilon_\tau,\kappa_\tau]$ and a contribution no bigger than a multiple of

$$\int_{\kappa_{C\tau}}^{\infty} [y^4\delta_{\tau/C}^2(y)+1]\varphi(y)\,dy \lesssim \int_{\kappa_{C\tau}}^{\infty} [C^2\tau^{-2}y^8e^{-3y^2/2}+e^{-y^2/2}]\,dy \lesssim \tau\kappa_\tau$$

from the interval $[\kappa_\tau,\infty)$.

For the proof of the third assertion it suffices to bound the fourth cumulant of θ_i given (Y_i,τ), in view of the second assertion. For any function h we have

$$(\log h)'''' = \frac{h''''}{h} - 4\frac{h'''}{h}\frac{h'}{h} + 12\frac{h''}{h}\Big(\frac{h'}{h}\Big)^2 - 3\Big(\frac{h''}{h}\Big)^2 - 6\Big(\frac{h'}{h}\Big)^4.$$

Combined with the formulas for $I'_{-1/2}$ and $I''_{-1/2}$ given before as well as $I'''_{-1/2}(y) = y^3I_{5/2}(y)+3yI_{3/2}(y)$ and $I''''_{-1/2}(y) = y^4I_{7/2}(y)+6y^2I_{5/2}(y)+3I_{3/2}(y)$, we find that the fourth cumulant can be written in the form

$$\begin{aligned}&\frac{y^4I_{7/2}(y)+6y^2I_{5/2}(y)+3I_{3/2}(y)}{I_{-1/2}(y)} - 4\frac{y^3I_{5/2}(y)+3yI_{3/2}(y)}{I_{-1/2}(y)}\frac{yI_{1/2}(y)}{I_{-1/2}(y)}\\ &\quad+12\frac{y^2I_{3/2}(y)+I_{1/2}(y)}{I_{-1/2}(y)}\Big(\frac{yI_{1/2}(y)}{I_{-1/2}(y)}\Big)^2 - 3\Big(\frac{y^2I_{3/2}(y)+I_{1/2}(y)}{I_{-1/2}(y)}\Big)^2 - 6\Big(\frac{yI_{1/2}(y)}{I_{-1/2}(y)}\Big)^4.\end{aligned}$$

As before we expand these expressions with the help of Lemmas 3.30 and 3.31, and next integrate separately over $[0,1/\varepsilon_{C^{-1}\tau}]$, $[1/\varepsilon_{C\tau},2\kappa_{\tau/C}]$, and $[2\kappa_{C\tau},\infty)$. The first interval gives

a negligible contribution. Following from the inequality $I_{-1/2}(y) \geq I_k(y)$ for $k \geq -1/2$ and Lemma 3.29 one can obtain that the dominating term in the second interval is $C\tau y^2 e^{y/2}$. This leads to

$$\int_0^{\kappa_{\tau/C}} \sup_{t\in[C^{-1}\tau_n, C\tau_n]} ty^2 e^{y^2/2}\varphi(y)\,dy \lesssim C\tau \int_0^{\kappa_{\tau/C}} y^2\,dy \lesssim \tau\zeta_\tau^3$$

On the last interval

$$\int_{y\geq 2\kappa_{C\tau}} y^4 e^{-y^2/2} \lesssim \kappa_\tau^{11}\tau^4 = o(\tau\zeta_\tau^3).$$

□

Proof of Theorem 3.10

Proof. The posterior distribution of θ_i given (Y_i, τ, λ_i) is normal with mean and variance

$$\hat{\theta}_i(\tau,\lambda_i) := \mathbb{E}(\theta_i \mid Y_i, \tau, \lambda_i) = \frac{\lambda_i^2\tau^2}{1+\lambda_i^2\tau^2} Y_i,$$
$$r_i^2(\tau,\lambda_i) := \operatorname{var}(\theta_i \mid Y_i, \tau, \lambda_i) = \frac{\lambda_i^2\tau^2}{1+\lambda_i^2\tau^2}.$$

Furthermore, the posterior distribution of λ_i given (Y_i, τ) possesses density function given by

$$\pi(\lambda_i \mid Y_i, \tau) \propto e^{-\frac{Y_i^2}{2(1+\lambda_i^2\tau^2)}} (1+\tau^2\lambda_i^2)^{-1/2}(1+\lambda_i^2)^{-1}.$$

The parameter $\theta_{0,i}$ is contained in $C_{ni}(L,\tau)$ if and only if $|\theta_{0,i} - \hat{\theta}_i(\tau)| \leq L\hat{r}_i(\alpha,\tau)$. We show that this is true, or not, for $\theta_{0,i}$ belonging to the three regions separately for S, L and M.

Case S: proof of (3.8). If $i \in S$, then $|\theta_{0,i} - \hat{\theta}_i(\tau)| \leq k_S\tau + \tau|Y_i|e^{Y_i^2/2}$, by the triangle inequality and Lemma 3.16(iii). Below we show that $\hat{r}_i(\alpha,\tau) \geq \tau z_\alpha c$, with probability tending to one, for z_α the standard normal upper α-quantile and every $c < 1/2$. Hence $\theta_{0,i} \in C_{ni}(L,\tau)$ as soon as $|Y_i|e^{Y_i^2/2} \leq Lz_\alpha c - k_S$.

For $i \in S$ the variable $|Y_i|$ is stochastically bounded by $|\theta_{0,i}| + |\varepsilon_i| \leq k_S\tau + |\varepsilon_i|$. Since the variables $|\varepsilon_i|$ are i.i.d. with quantile function $u \mapsto \Phi^{-1}((u+1)/2) \leq \sqrt{2\log(2/(1-u))}$, a fraction $1-\gamma$ of the variables Y_i with $i \in S$ is bounded above by $k_S\tau + \sqrt{2\log(2/\gamma)} + \delta = k_S\tau + \zeta_{\gamma/2} + \delta$, with probability tending to 1, for any $\delta > 0$. Then the corresponding fraction of parameters $\theta_{0,i}$ is contained in their credible interval if L is chosen big enough that

$$Lz_\alpha c - k_S \geq (k_S\tau + \zeta_{\gamma/2} + \delta)e^{(k_S\tau+\zeta_{\gamma/2}+\delta)^2/2} \leq \frac{2}{\gamma}\zeta_{\gamma/2}(1+\varepsilon),$$

where $\varepsilon \to 0$ if $\gamma \to 0$ and can chosen arbitrarily small if δ is chosen small and $\tau \to 0$. This is certainly true for L_S as in the theorem.

We finish by proving the lower bound for the radius $\hat{r}_i(\alpha,\tau)$. Because the conditional distribution of θ_i given (Y_i, τ, λ_i) is normal with mean $\hat{\theta}_i(\tau,\lambda_i)$ it follows by Anderson's

lemma that $\Pi\big(\theta_i : |\theta_i - \hat\theta_i(\tau)| > r \,|\, Y_i, \tau, \lambda_i\big) \geq \Pi\big(\theta_i : |\theta_i - \hat\theta_i(\tau, \lambda_i)| > r \,|\, Y_i, \tau, \lambda_i\big)$, for any $r > 0$. Furthermore, by the monotonicity of the variance in λ_i of this conditional distribution, the last function is increasing in λ_i. If $\tilde\pi(\cdot \,|\, \tau)$ is the probability density given by

$$\tilde\pi(\lambda_i \,|\, \tau) \propto (\lambda_i^2\tau^2 + 1)^{-1/2}(1 + \lambda_i^2)^{-1},$$

then $\lambda_i \mapsto \pi(\lambda_i \,|\, Y_t, \tau)/\tilde\pi(\lambda_i \,|\, \tau)$ is increasing. Combining the preceding observations with Lemma 3.18, we see that

$$\begin{aligned}\alpha &= \int_0^\infty \Pi(\theta_i : |\theta_i - \hat\theta_i(\tau)| > \hat r_i(\alpha, \tau) \,|\, Y_i, \tau, \lambda_i)\pi(\lambda_i \,|\, Y_i, \tau)\, d\lambda_i \\ &\geq \int_0^\infty \Pi(\theta_i : |\theta_i - \hat\theta_i(\tau, \lambda_i)| > \hat r_i(\alpha, \tau) \,|\, Y_i, \tau, \lambda_i)\tilde\pi(\lambda_i \,|\, \tau)\, d\lambda_i. \end{aligned} \tag{3.30}$$

On the other hand, since $\mathrm{sd}(\theta_i \,|\, Y_i, \tau, \lambda_i) \geq \tau/2(1 + o(1))$, for $\lambda_i \geq 1/2$, the normality of the conditional distribution of θ_i given (Y_i, τ, λ_i) gives that

$$\begin{aligned}&\int_0^\infty \Pi\Big(\theta_i : |\theta_i - \hat\theta_i(\tau, \lambda_i)| > z_\alpha\tau/2(1 + o(1))) \,|\, Y_i, \tau, \lambda_i\Big)\tilde\pi(\lambda_i \,|\, \tau)\, d\lambda_i \\ &\qquad \geq 2\alpha\,\tilde\Pi(\lambda_i \geq 1/2 \,|\, \tau) \geq 2\alpha \times 2/3 > \alpha. \end{aligned} \tag{3.31}$$

Here the second last inequality follows from

$$\frac{\int_0^{1/2}(\lambda_i^2\tau^2 + 1)^{-1/2}\,(1 + \lambda_i^2)^{-1}\, d\lambda}{\int_0^\infty(\lambda_i^2\tau^2 + 1)^{-1/2}\,(1 + \lambda_i^2)^{-1}\, d\lambda_i} \to \frac{\int_0^{1/2}(1 + \lambda_i^2)^{-1}\, d\lambda_i}{\int_0^\infty(1 + \lambda_i^2)^{-1}\, d\lambda_i} < \frac{1}{3},$$

as $\tau \to 0$, by two applications of the dominated convergence theorem. Combination of (3.30) and (3.31) shows that $\hat r_i(\alpha, \tau) \geq z_\alpha\tau/2(1 + o(1))$.

Case L: proof of (3.10). If $i \in L$, then $|\theta_{0,i} - \hat\theta_i(\tau)| \leq |\theta_{0,i} - Y_i| + |Y_i - \hat\theta_i(\tau)| \leq |\varepsilon_i| + 2\zeta_\tau^{-1}$, eventually, provided $|Y_i| \geq A\zeta_\tau$ for some constant $A > 1$, by the triangle inequality and Lemma 3.16(i). Below we show that $\hat r_i(\alpha, \tau) \geq z_\alpha + o(1)$, with probability tending to one. It then follows that $\theta_{0,i} \in C_{ni}(L, \tau)$ as soon as $|Y_i| \geq A\zeta_\tau$ and $|\varepsilon_i| \leq Lz_\alpha + o(1) - 2\zeta_\tau^{-1} = Lz_\alpha + o(1)$.

For $i \in L$ the variable $|Y_i|$ is lower bounded by $|\theta_{0,i}| - |\varepsilon_i| \geq k_L\zeta_\tau - |\varepsilon_i|$ and hence $|Y_i| \geq A\zeta_\tau$ if $|\varepsilon_i| \leq (k_L - A)\zeta_\tau$. This is automatically satisfied if $|\varepsilon_i| \leq Lz_\alpha + o(1)$, for constants L with $L \ll \zeta_\tau$. As for the proof of Case S we have that $|\varepsilon_i| \leq Lz_\alpha + o(1)$ with probability tending to one for a fraction γ of the indices $i \in S$ if $L \geq z_\alpha^{-1}\zeta_{\gamma/2} + \delta$, for some $\delta > 0$.

The proof that $\hat r_i(\alpha, \tau) \geq z_\alpha + o(1)$ follows the same lines as the proof of the corresponding result in Case S, expressed in (3.30) and (3.31), but with the true density π instead of $\tilde\pi$. Inequality (3.30) with π instead of $\tilde\pi$ is valid by Anderson's lemma, while in (3.31) we replace $z_\alpha\tau/2(1 + o(1))$ by $z_\alpha + o(1)$. Since $\mathrm{var}(\theta_i \,|\, Y_i, \tau, \lambda_i) \geq g_\tau/(1 + g_\tau) = 1 + o(1)$ for every $\lambda_i \geq g_\tau/\tau$ and $g_\tau \to \infty$, the desired result follows if $\Pi\big(\lambda_i \geq g_\tau/\tau \,|\, Y_i, \tau\big)$ is eventually bigger than $2/3$, for every i such that $|Y_i| \geq A\zeta_\tau$. Now by the form of $\pi(\lambda_i \,|\, Y_i, \tau)$, for

any $c, d > 0$,

$$\Pi(\lambda_i \le g_\tau/\tau \mid Y_i, \tau) \le \frac{e^{-\frac{Y_i^2}{2(1+c^2)}} \int_0^{c/\tau} (1+\lambda^2)^{-1}\, d\lambda + e^{-\frac{Y_i^2}{2(1+g_\tau^2)}} \int_{c/\tau}^{g_\tau/\tau} (1 + c^2/\tau^2)^{-1}\, d\lambda}{e^{-\frac{Y_i^2}{2(1+d^2 g_\tau^2)}} \int_{dg_\tau/\tau}^{2dg_\tau/\tau} (1 + 4d^2 g_\tau^2)^{-1/2} (1 + 4d^2 g_\tau^2/\tau^2)^{-1}\, d\lambda}$$
$$\lesssim \frac{\exp\left[-\frac{Y_i^2}{2}\left(\frac{1}{1+c^2} - \frac{1}{1+d^2 g_\tau^2}\right)\right] + \exp\left[-\frac{Y_i^2}{2}\left(\frac{1}{1+g_\tau^2} - \frac{1}{1+d^2 g_\tau^2}\right)\right] g_\tau \tau}{(g_\tau/\tau)(1/g_\tau)(\tau^2/g_\tau^2)}.$$

For $|Y_i| > A\zeta_\tau$ and $A > 1$ we can choose c sufficiently close to zero so that the first exponential is of order $\tau^{A'}$ for some $A' > 1$. Then it is much smaller than the denominator, which is of order τ/g_τ^2, provided g_τ tends to infinity slowly. If we choose $d > 1$, then the term involving the second exponential will also tend to zero for $|Y_i| > A\zeta_\tau$ as soon as $e^{-c\zeta_\tau^2/g_\tau^2} g_\tau^3 \to 0$, for a sufficiently small constant c. This is true (for any $c > 0$) for instance if $g_\tau = \sqrt{\zeta_\tau}$. Then the quotient tends to zero, and is certainly smaller than $1/3$.

Case M: proof of (3.9). We show below that $\hat{r}_i(\alpha, \tau) \lesssim U_\tau := \tau(1 \vee |Y_i| e^{Y_i^2/2})$, with probability tending to one, whenever $i \in M$. By Lemma 3.16(iii) exactly the same bound is valid for $|\hat{\theta}_i(\tau)|$. If $|\hat{\theta}_i(\tau)| + \hat{r}_i(\alpha, \tau) \lesssim U_\tau$, but $|\theta_{0,i}| \gg U_\tau$ then $\theta_{0,i} \notin C_{ni}(L, \tau)$ eventually, and hence it suffices to prove that the probability of the event that $|\theta_{0,i}| \gg U_\tau$ tends to one whenever $i \in M$. Consider two cases. If $|\theta_{0,i}| \le 1$, then $|Y_i| \le 1 + |\varepsilon_i| = O_P(1)$ and hence $U_\tau = O_P(\tau)$. For $i \in M$, we have $|\theta_{0,i}| \gg \tau$ and hence $|\theta_{0,i}| \gg U_\tau$ with probability tending to one. On the other hand, if $|\theta_{0,i}| \ge 1$ but $|\theta_{0,i}| \le k_M \zeta_\tau$, then $|Y_i| \le k\zeta_\tau$ with probability tending to one for any $k > k_M$, and hence $U_\tau \lesssim \tau\zeta_\tau e^{k^2\zeta_\tau^2/2} = \tau^{1-k^2}\zeta_\tau$. Since $k_M < 1$ we can choose $k < 1$, so that $\tau^{1-k^2}\zeta_\tau \to 0$, and again we have $|\theta_{0,i}| \gg U_\tau$ with probability tending to one.

We finish by proving that $\hat{r}_i(\alpha, \tau) \lesssim U_\tau$, with probability tending to one. As a first step we show that, for $k < 1$,

$$\lim_{M\to\infty} \sup_{|y| \le k\zeta_\tau} \Pi(\lambda_i \ge M \mid Y_i = y, \tau) \to 0. \tag{3.32}$$

By the explicit form of the posterior density of λ_i we have

$$\Pi(\lambda_i \ge M \mid Y_i = y, \tau) \le \frac{\int_M^\infty e^{-\frac{y^2}{2(1+\lambda_i^2\tau^2)}} (1+\lambda_i^2\tau^2)^{-1/2} (1+\lambda_i^2)^{-1}\, d\lambda_i}{\int_1^2 e^{-\frac{y^2}{2(1+\lambda_i^2\tau^2)}} (1+\lambda_i^2\tau^2)^{-1/2} (1+\lambda_i^2)^{-1}\, d\lambda_i}$$
$$\le e^{y^2/2} 5\sqrt{2} \int_M^\infty e^{-\frac{y^2}{2(1+\lambda_i^2\tau^2)}} (1+\lambda_i^2\tau^2)^{-1/2} (1+\lambda_i^2)^{-1}\, d\lambda_i.$$

We split the remaining integral over the intervals $[M, \tau^{-a})$ and $[\tau^{-a}, \infty)$, for some $a < 1$. On the first interval we use that $y^2/(1+\lambda_i^2\tau^2) = y^2 + o(1)$, uniformly in $|y| \lesssim \zeta_\tau$ and $\lambda_i \le \tau^{-a}$, while on the second we simply bound the factor $e^{-y^2/(2(1+\lambda_i^2\tau^2))}$ by 1, to see that the preceding display is bounded above by

$$e^{y^2/2} 5\sqrt{2}\left[e^{-y^2/2} e^{o(1)} \int_M^{\tau^{-a}} (1+\lambda_i^2)^{-1}\, d\lambda_i + \int_{\tau^{-a}}^\infty (1+\lambda_i^2)^{-1}\, d\lambda_i\right].$$

The first term in square brackets (times the leading term) contributes less than a multiple of $\int_M^\infty \lambda^{-2}\,d\lambda = 1/M$, while the second term contributes less than $e^{y^2/2}\tau^a \leq \tau^{-k^2+a}$, for $|y| \leq k\zeta_\tau$, which tends to zero if $a > k^2$. This concludes the proof of (3.32).

By the triangle inequality, for any $M > 0$,

$$\int_0^\infty \Pi(\theta_i : |\theta_i - \hat\theta_i(\tau)| \geq r + |\hat\theta_i(\tau,\lambda_i) - \hat\theta_i(\tau)| \mid Y_i,\lambda_i,\tau)\pi(\lambda_i \mid Y_i,\tau)\,d\lambda_i$$
$$\leq \int_0^M \Pi(\theta_i : |\theta_i - \hat\theta_i(\tau,\lambda_i)| \geq r \mid Y_i,\lambda_i,\tau)\pi(\lambda_i \mid Y_i,\tau)\,d\lambda_i + \Pi(\lambda_i \geq M \mid Y_i,\tau).$$

For sufficiently large M the second term on the far right is smaller than $\alpha/2$ by the preceding paragraph and for $r = z_{\alpha/4}\sup_{\lambda\leq M} r_i(\tau,\lambda)$ the first term on the right is smaller than $\alpha/2$ as well, by the normality of θ_i given (Y_i,λ_i,τ) and the definition of $r_i(\tau,\lambda_i)$. The inequality remains valid if $|\hat\theta_i(\tau,\lambda_i) - \hat\theta_i(\tau)|$ in the first line is replaced by $\sup_{\lambda\leq M}|\hat\theta_i(\tau,\lambda_i)| + |\hat\theta_i(\tau)|$. It follows that

$$\hat r_i(\alpha,\tau) \leq z_{\alpha/4}\sup_{\lambda\leq M} r_i(\tau,\lambda) + \sup_{\lambda\leq M}|\hat\theta_i(\tau,\lambda_i)| + |\hat\theta_i(\tau)|.$$

The first term is bounded above by $M\tau$, and the second by $M\tau|Y_i|$, by the definitions of $r_i(\tau,\lambda)$ and $\hat\theta_i(\tau,\lambda)$, while $|\hat\theta_i(\tau)| \leq \tau|Y_i|e^{Y_i^2/2}$, by Lemma 3.16(iii). This concludes the proof that $\hat r_i(\alpha,\tau) \lesssim U_\tau$. □

Lemma 3.18. If $f_1, f_2 : [0,\infty) \to [0,\infty)$ are probability densities such that f_2/f_1 is monotonely increasing, then, for any monotonely increasing function h,

$$\mathbb{E}_{f_1} h(X) \leq \mathbb{E}_{f_2} h(X).$$

Proof. Define $g = f_2/f_1$. Since $\int_0^\infty f_1(x)dx = \int_0^\infty f_1(x)g(x)\,dx$ and g is monotonely increasing, there exists an $x_0 > 0$ such that $g(x) \leq 1$ for $x < x_0$ and $g(x) \geq 1$ for $x > x_0$. Therefore

$$0 = h(x_0)\int_0^\infty f_1(x)\Big(g(x) - 1\Big)\,dx$$
$$\leq \int_0^{x_0} f_1(x)h(x)\Big(g(x) - 1\Big)\,dx + \int_{x_0}^\infty f_1(x)h(x)\Big(g(x) - 1\Big)\,dx.$$

By the definition of g the right side is $\mathbb{E}_{f_2}h(X) - \mathbb{E}_{f_1}h(X)$. □

3.6.5 Proofs for the adaptive credible sets

Proof of Theorem 3.13

Proof. To simplify notation set $T_n = [C^{-1}\tilde\tau_n, C\tilde\tau_n]$, where $\tilde\tau_n = \tau_n(\tilde p_n)$.

First we deal with the empirical Bayes credible sets. Since $\hat\tau_n \in T_n$ with probability tending to one by Condition 8,

$$P_{\theta_0}\Big(\theta_0 \notin \hat C_n(\hat\tau_n, L)\Big) = P_{\theta_0}\Big(\|\theta_0 - \hat\theta(\hat\tau_n)\|_2 > L\hat r(\alpha,\hat\tau_n)\Big)$$

$$\leq P_{\theta_0}\Big(\sup_{\tau\in T_n} \|\theta_0 - \hat\theta(\tau)\|_2 > L \inf_{\tau\in T_n} \hat r(\alpha,\tau)\Big) + o(1).$$

By Lemma 3.19 $\inf_{\tau\in T_n} \hat r(\alpha,\tau) \gtrsim \sqrt{n\tilde\tau_n\zeta_{\tilde\tau_n}}$, with probability tending to one. Therefore it suffices to show that $\sup_{\tau\in T_n} \|\theta_0 - \hat\theta(\tau)\|_2 = O_P(\sqrt{n\tilde\tau_n\zeta_{\tilde\tau_n}})$. We show this by bounding the second moment of this variable.

We split the sum in $\|\hat\theta(\tau) - \theta_0\|_2^2 = \sum_i(\hat\theta_i(\tau) - \theta_{0,i})^2$ in two parts, according to the values of $\theta_{0,i}$. Set $\tilde\zeta = A\sqrt{2\log(n/q)}$, for q as in (3.14).

If $|\theta_{0,i}| \geq \zeta_{\tilde\tau_n}/5$, then we first use Lemma 3.16(ii) together with the triangle inequality to see that $|\hat\theta_i(\tau) - \theta_{0,i}| \lesssim \zeta_\tau + |Y_i - \theta_{0,i}|$, as $\tau \to 0$, whence

$$\mathbb{E}_{\theta_{0,i}} \sup_{\tau\in T_n} (\theta_{0,i} - \hat\theta_i(\tau))^2 \lesssim \sup_{\tau\in T_n} \zeta_\tau^2 + \operatorname{var}_{\theta_{0,i}} Y_i \lesssim \zeta_{\tilde\tau_n}^2.$$

By the excessive-bias restriction

$$\frac{\zeta_{\tilde\tau_n}^2}{25}\Big|\Big\{i : \frac{\zeta_{\tilde\tau_n}}{5} < |\theta_{0,i}| < \tilde\zeta\Big\}\Big| \leq \sum_{i:|\theta_{0,i}|\leq\tilde\zeta} \theta_{0,i}^2 \lesssim q\log(n/q) \lesssim \tilde p\log(ne/(C_s\tilde p)).$$

Since $\log(ne/(C_s\tilde p))/\zeta_{\tilde\tau_n}^2 \to 1$, it follows that there are fewer than a constant times $\tilde p$ parameters with $|\theta_{0,i}| \geq \zeta_{\tilde\tau_n}/5$ and hence their total contribution to the sum is bounded by $\tilde p\zeta_{\tilde\tau_n}^2$.

For parameters such that $|\theta_{0,i}| \leq \zeta_{\tilde\tau_n}/5$ we use the triangle inequality $|\hat\theta_i(\tau) - \theta_{0,i}| \leq |\hat\theta_i(\tau)| + |\theta_{0,i}|$, and next further bound $|\hat\theta_i(\tau)|$ by $\tau|Y_i|e^{Y_i^2/2}$ in case $|Y_i - \theta_{0,i}| \leq \zeta_{\tilde\tau_n}$, which is valid in view of Lemma 3.16 (iii), and further bound $|\hat\theta_i(\tau)| \leq |Y_i|$ by $|Y_i - \theta_{0,i}| + |\theta_{0,i}|$, otherwise. This gives

$$\begin{aligned}\mathbb{E}_{\theta_{0,i}} \sup_{\tau\in T_n} |\theta_{0,i} - \hat\theta_i(\tau)|^2 \lesssim{}& \mathbb{E}_{\theta_{0,i}} \tilde\tau_n^2|Y_i|^2 e^{Y_i^2}\mathbf{1}_{|Y_i-\theta_{0,i}|\leq\zeta_{\tilde\tau_n}} \\ &+ \mathbb{E}_{\theta_{0,i}}|Y_i - \theta_{0,i}|^2\mathbf{1}_{|Y_i-\theta_{0,i}|>\zeta_{\tilde\tau_n}} + \theta_{0,i}^2.\end{aligned}$$

The second expectation on the right is bounded above by $\tilde\tau_n\zeta_{\tilde\tau_n}$. The first expectation on the right is equal to $\tau^2\int_{-\zeta_\tau}^{\zeta_\tau}(y+\theta)^2e^{(y+\theta)^2}\varphi(y)\,dy \lesssim \tau^2\zeta_\tau^2\int_0^{\zeta_\tau} e^{\theta^2+2y|\theta|}e^{y^2/2}\,dy$, for $\tau = \tilde\tau_n$ and $\theta = \theta_{0,i}$. For $|\theta| \lesssim \zeta_\tau^{-1}$, the exponential factor $e^{\theta^2+2y|\theta|}$ is uniformly bounded, and the whole expression is bounded by a multiple of $\tau^2\zeta_\tau^2\int_0^{\zeta_\tau} e^{y^2/2}\,dy \lesssim \tau\zeta_\tau$. For $|\theta| \gtrsim \zeta_\tau^{-1}$, but $|\theta| \leq \zeta_\tau/5$, the exponential factor is bounded above by $e^{\zeta_\tau^2/25+2\zeta_\tau^2/5} = \tau^{-22/25}$ and the whole expression is bounded above by $\tau^{3/22}\zeta_\tau \lesssim \theta^2$. Thus in both cases the first equation is bounded above by a multiple of $\tilde\tau_n\zeta_{\tilde\tau_n} + \theta_{0,i}^2$.

Combining the above two cases we find

$$\mathbb{E}_{\theta_0} \sup_{\tau\in T_n} \|\hat\theta(\tau) - \theta_0\|_2^2 \lesssim \tilde p\zeta_{\tilde\tau_n}^2 + n\tilde\tau_n\zeta_{\tilde\tau_n} + \sum_{i:|\theta_{0,i}|<\zeta_{\tilde\tau_n}/5} \theta_{0,i}^2.$$

Since $\zeta_{\tilde\tau_n}^2 \sim \log(n/\tilde p) \leq \log(ne/(C_sq)) \sim \log(n/q)$ and $1/5 < 1$, the last term is bounded above by a multiple of $q\log(n/q) \lesssim \tilde p\log(n/\tilde p)$ by the excessive-bias restriction, whence

the whole expression is bounded above $n\tilde{\tau}_n\zeta_{\tilde{\tau}_n} \asymp \tilde{p}\log(n/\tilde{p})$. This concludes the proof of the coverage of the empirical Bayes credible balls.

The proof of their rate-adaptive size follows along the same lines.

Next we deal with the hierarchical Bayes credible sets. By Lemma 3.20 and the triangle inequality

$$\begin{aligned} P_{\theta_0}\Big(\theta_0 \notin \hat{C}_n(L)\Big) &\leq P_{\theta_0}\Big(\|\theta_0 - \hat{\theta}\|_2 > L\hat{r}(\alpha)\Big) \\ &\leq P_{\theta_0}\Big(\|\theta_0 - \hat{\theta}(\tilde{\tau}_n)\|_2 + \|\hat{\theta} - \hat{\theta}(\tilde{\tau}_n)\|_2 > LA\sqrt{n\zeta_{\tilde{\tau}_n}\tilde{\tau}_n}\Big) + o(1). \end{aligned}$$

The proof for the empirical Bayes set as just given shows that $\|\theta_0 - \hat{\theta}(\tilde{\tau}_n)\|_2 = O_P(\sqrt{n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}})$. Therefore, it is sufficient to show that $\|\hat{\theta} - \hat{\theta}(\tau_n)\|_2 = O_P(\sqrt{n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}})$. Since $\hat{\theta} = \int \hat{\theta}(\tau)\,\pi(\tau \mid Y_i)\,d\tau$, Jensen's inequality gives

$$\begin{aligned} \|\hat{\theta} - \hat{\theta}(\tilde{\tau}_n)\|_2^2 &\leq \int_{1/n}^1 \|\hat{\theta}(\tau) - \hat{\theta}(\tilde{\tau}_n)\|_2^2 \pi(\tau \mid Y^n)\,d\tau \\ &\leq \sup_{\tau\in T_n} \|\hat{\theta}(\tau) - \hat{\theta}(\tilde{\tau}_n)\|_2^2 + \sup_{\tau\in[1/n,1]} \|\hat{\theta}(\tau) - \hat{\theta}(\tilde{\tau}_n)\|_2^2\, \Pi(\tau \notin T_n \mid Y^n). \end{aligned} \tag{3.33}$$

The first term on the right hand side is bounded from above by $4\sup_{\tau\in T_n}\|\hat{\theta}(\tau) - \theta_0\|_2^2$, and was already seen to be $O_P(n\tilde{\tau}_n\zeta_{\tilde{\tau}_n})$. By the triangle inequality and Lemma 3.16 (i)+(ii) the second supremum on the right hand side is bounded by

$$4 \sup_{\tau\in[1/n,1]} \|\hat{\theta}(\tau) - Y^n\|_2^2 \leq 4n \sup_{\tau\in[1/n,1]} \zeta_\tau^2 \lesssim n\log n.$$

By Lemma 3.21 we can choose the constant C in the definition of T_n such that $\Pi(\tau \notin T_n \mid Y^n) \leq e^{-c_3\tilde{p}}$, for a constant $c_3 > 0$. For $\tilde{p} \geq (2/c_3)\log n$ the probability $\Pi(\tau \notin T_n \mid Y^n)$ is of the order n^{-2}, and the second term on the right hand side of (3.33) is negligible. □

Proof of Theorem 3.14

Proof. The proof for the empirical Bayes procedure closely follows the proof of Theorem 3.10. The lower bounds $\hat{r}_i(\alpha,\tau) \geq \tau z_\alpha(1 + o(1))$ and $\hat{r}_i(\alpha,\tau) \geq z_\alpha + o(1)$ in the cases S and L, and the upper bound $\hat{r}_i(\alpha,\tau) \leq \tau(1 \vee |Y_i|e^{Y_i^2/2})$ in case M, with probability tending to one, remain valid when τ is replaced by $\hat{\tau}_n$. The remainders of the arguments then go through with minor changes, where it is used that $\hat{\tau}_n \geq 1/n$, $\zeta_{\hat{\tau}_n} \leq \sqrt{2\log n}$ and $\hat{\tau}_n \leq \tau_n(p)$ with probability tending to one by Condition 4. Note the slightly changed right boundary of the set S_a and left boundary of the set L_a, which refer to "extreme" cases.

In the proof for the hierarchical Bayes method, we denote by $\hat{\theta}_i$ the ith coordinate of the hierarchical posterior mean $\hat{\theta}$ and by $\hat{r}_i(\alpha)$ the (Bayesian) radius of the marginal hierarchical Bayes credible interval. Hence $\theta_{0,i}$ is contained in this credible interval if $|\theta_{0,i} - \hat{\theta}_i| \leq L\hat{r}_i(\alpha)$.

By Lemma 3.6 we have that $\Pi(1/n < \tau < 5t_n \mid Y^n) \to 1$ under Condition 6, or $\Pi(1/n < \tau < (\log n)t_n \mid Y^n) \to 1$ under the weaker Condition 7.

Case S_a: proof of the hierarchical Bayes version of (3.15). For $i \in S_a$ we have $|Y_i| \leq k_S/n + |\varepsilon_i|$. Because the $1-\gamma$-quantile of the absolute errors $|\varepsilon_i|$ is bounded above by $\zeta_{\gamma/2}$,

the set S_γ of coordinates $i \in S_a$ such that $|Y_i| \le \zeta_{\gamma/2} + \delta$ contains at least a fraction $1-\gamma$ of the elements of S_a, with probability tending to one. We show below that with probability tending to one both $\hat{r}_i(\alpha) \ge c|\hat{\theta}_i|z_{\alpha/2}\zeta_{\gamma/2}$ and $\hat{r}_i(\alpha) \ge z_\alpha/(2n)$ for $i \in S_\gamma$, and any $c < 1/2$. Then $|\hat{\theta}_i - \theta_{0,i}| \le |\hat{\theta}_i| + k_S/n \le [(cz_{\alpha/2}\zeta_{\gamma/2})^{-1} + (2/z_\alpha)k_S]\hat{r}_i(\alpha)$, and hence $\theta_{0,i}$ is contained in its credible interval for every $i \in S_\gamma$ if $L \ge (cz_{\alpha/2}\zeta_{\gamma/2})^{-1} + (2/z_\alpha)k_S$.

To show that $\hat{r}_i(\alpha) \ge c|\hat{\theta}_i|z_{\alpha/2}\zeta_{\gamma/2}$ for $i \in S_\gamma$, we assume $Y_i > 0$ for simplicity. Then $\hat{\theta}_i(\tau,\lambda_i) > 0$ for every (τ,λ_i) and hence so is $\hat{\theta}_i$. By its definition $\hat{\theta}_i(\tau,\lambda_i) = r_i^2(\tau,\lambda_i)Y_i$. Since $r_i(\tau,\lambda_i) \le 1$, it follows that $\hat{\theta}_i(\tau,\lambda_i) \le r_i(\tau,\lambda_i)(\zeta_{\gamma/2} + \delta)$, for every $i \in S_\gamma$. If $\hat{\theta}_i(\tau,\lambda_i) \ge \hat{\theta}_i/2$, then $r_i(\tau,\lambda_i) \ge \hat{\tau}_i(2\zeta_{\gamma/2} + 2\delta)$ and we can conclude, using Anderson's lemma and the conditional normal distribution of θ_i given (Y_i,λ_i,τ) with variance $r_i^2(\tau,\lambda_i)$, that $\Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge z_{\alpha/2}\hat{\theta}_i/(2\zeta_{\gamma/2} + 2\delta) \mid Y_i,\tau,\lambda_i\Big) \ge \alpha$. If $\hat{\theta}_i(\tau,\lambda_i) \le \hat{\theta}_i/2$, then $\theta_i \le \hat{\theta}_i(\tau,\lambda_i)$ implies that $|\theta_i - \hat{\theta}_i| \ge \hat{\theta}_i/2$, and hence $\Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge \hat{\theta}_i/2 \mid Y_i,\tau,\lambda_i\Big) \ge \Pi\Big(\theta_i : \theta_i \le \hat{\theta}_i(\tau,\lambda_i) \mid Y_i,\tau,\lambda_i\Big) = 1/2$, since $\hat{\theta}_i(\tau,\lambda_i)$ is the median of the conditional normal distribution of θ_i. For $c_0 = (1/2) \wedge (z_{\alpha/2}/(2\zeta_{\gamma/2} + 2\delta))$ and $\alpha \le 1/2$, we have that $\Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge c\hat{\theta}_i\Big) \ge \alpha$ in both cases, and hence

$$\Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge c_0\hat{\theta}_i \mid Y^n\Big)$$
$$\int_{1/n}^1 \int_0^\infty \Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge c_0\hat{\theta}_i \mid Y_i,\tau,\lambda_i\Big)\pi(\lambda_i \mid Y_i,\tau)\pi(\tau \mid Y^n)\, d\lambda_i\, d\tau \ge \alpha.$$

Thus $\hat{r}_i(\alpha) \ge c_0\hat{\theta}_i$ by the definition of $\hat{r}_i(\alpha)$.

For the proof that $\hat{r}_i(\alpha) \ge z_\alpha/(2n)$, we first note that, similarly to (3.30),

$$\alpha = \Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge \hat{r}_i(\alpha) \mid Y^n\Big)$$
$$\ge \int_{1/n}^1 \int_0^\infty \Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i(\tau,\lambda_i)| \ge \hat{r}_i(\alpha) \mid Y_i,\tau,\lambda_i)\pi(\lambda_i \mid Y_i,\tau\Big)\pi(\tau \mid Y^n)\, d\lambda_i\, d\tau$$

On the other hand, since $r_i(\tau,\lambda_i) \ge 1/(2n)(1+o(1))$, whenever $\tau \in [1/n, 5t_n]$ and $\lambda_i > 1/2$, we have similarly to (3.31),

$$\int_{1/n}^1 \int_0^\infty \Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge z_\alpha/(2n) \mid Y_i,\tau,\lambda_i\Big)\pi(\lambda_i \mid Y_i,\tau)\pi(\tau \mid Y^n)\, d\lambda_i\, d\tau$$
$$\ge \int_{1/n}^{5t_n} \int_{1/2}^\infty \Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge z_\alpha r_i(\tau,\lambda_i) \mid Y_i,\tau,\lambda_i\Big)\pi(\lambda_i \mid Y_i,\tau)\pi(\tau \mid Y^n)\, d\lambda_i\, d\tau$$
$$\ge \int_{1/n}^{5t_n} (4\alpha/3)\pi(\tau \mid Y^n) d\tau > \alpha,$$

where the lower bound $4\alpha/3$ follows as in (3.31). Together the two preceding displays imply that $\hat{r}_i(\alpha) \ge z_\alpha/(2n)$.

Case L_a: proof of the hierarchical Bayes version of(3.17). If $i \in L$, then $|\theta_{0,i}| \ge k_L\sqrt{2\log n} = k_L\zeta_{1/n}$ and hence $|Y_i| \ge k_L\zeta_{1/n} - |\varepsilon_i|$. The subset L_γ of i with $|\varepsilon_i| \le \zeta_{\gamma/2} + \delta$ contains a fraction of at least $1-\gamma$ of the elements of L_a eventually with probability tending to one, and $|Y_i| \ge k\zeta_{1/n}$ for every $i \in L_\gamma$ and some constant $k > 1$. Then $|Y_i - \theta_i(\tau)| \lesssim \zeta_\tau^{-1}$

for $\tau \to 0$ and $|Y_i - \theta_i(\tau)| \lesssim (\log \zeta_{1/n})/\zeta_{1/n}$ for τ bounded away from zero, by Lemma 3.16 (i) and (vii), respectively, and hence $|Y_i - \hat{\theta}_i|$ tends to zero, by Jensen's inequality. It follows that $|\theta_{0,i} - \hat{\theta}_i| \le |\theta_{0,i} - Y_i| + |Y_i - \hat{\theta}_i| \le \zeta_{\gamma/2} + \delta'$ for ever $i \in L_\gamma$ with probability tending to one. We can prove that $\hat{r}_i(\alpha) \ge z_\alpha(1 + o(1))$ similarly as in the proof for Case L in the proof of Theorem 3.10 (adapted similarly as in the proof for case S_a), but now using that $r_i(\tau, \lambda_i) \ge 1 + o(1)$, whenever $\tau \in [1/n, 5t_n]$ and $\lambda_i \ge g_\tau/\tau$, for some $g_\tau \to \infty$. Thus $|\theta_{0,i} - \hat{\theta}_i| \le L\hat{r}_i(\alpha)$ with probability tending to one, if $Lz_\alpha \ge \zeta_{\gamma/2} + \delta'$.

Case M_a: proof of the hierarchical Bayes version of (3.16). First assume that Condition 6 holds, so that $\Pi(\tau \le 5t_n \mid Y^n) \to 1$ in probability, by Lemma 3.6, and in fact $\Pi(\tau \le 5t_n \mid Y^n) \le e^{-c_0 p_n}$, for some $c_0 > 0$ by the proof of the lemma. Since $i \in M_a$ we have that $|Y_i| \le |\theta_{0,i}| + |\varepsilon_i| \le k\zeta_{\tau_n}$, with probability tending to one and some $k < 1$. We show below that both $\hat{r}_i(\alpha)$ and $|\hat{\theta}_i|$ are bounded above by $t_n(1 \vee |Y_i|)e^{Y_i^2/2}$, with probability tending to one. The argument as in the proof Theorem 3.10, split in the cases that $|\theta_{0,i}|$ is smaller or bigger than 1, then goes through and shows that $\theta_{0,i}$ is not contained in the credible interval, with probability tending to one.

By the triangle inequality, for any $r > 0$,

$$\Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i| \ge r + |\hat{\theta}_i(\tau, \lambda_i) - \hat{\theta}_i| \mid Y_i, \lambda_i, \tau\Big) \le \Pi\Big(\theta_i : |\theta_i - \hat{\theta}_i(\tau, \lambda_i)| \ge r \mid Y_i, \lambda_i, \tau\Big).$$

For $r \ge z_{\alpha/4} r_i(\tau, \lambda_i)$ the right side is at most $\alpha/2$. For given M define

$$r_i := z_{\alpha/4} \sup_{\substack{\tau \in [1/n, 5t_n] \\ \lambda_i \le M}} r_i(\tau, \lambda_i) + \sup_{\substack{\tau \in [1/n, 5t_n] \\ \lambda_i \le M}} |\hat{\theta}_i(\tau, \lambda_i)| + |\hat{\theta}_i|.$$

Then it follows that

$$\begin{aligned} &\int_{1/n}^1 \int_0^\infty \Pi(\theta_i : |\theta_i - \hat{\theta}_i| \ge r_i \mid Y_i, \lambda_i, \tau) \pi(\lambda_i \mid Y_i, \tau) \pi(\tau \mid Y^n)\, d\lambda_i\, d\tau \\ &\quad \le \alpha/2 + \int_{1/n}^{5t_n} \int_M^\infty \pi(\lambda_i \mid \tau, Y_i) \pi(\tau \mid Y^n)\, d\lambda_i\, d\tau + \int_{5t_n}^1 \pi(\tau \mid Y^n)\, d\tau. \end{aligned}$$

By (3.32) the second term on the right can be made arbitrarily small by choosing large M, and the third term tends to zero by Lemma 3.6. We conclude that the left side is then smaller than α which implies that $\hat{r}_i(\alpha) \le r_i$. Now by the definitions of $r_i(\tau, \lambda_i)$ and $\hat{\theta}_i(\tau, \lambda_i)$ the suprema in the definition of r_i are bounded by $z_{\alpha/4} M 5t_n$ and $M 5t_n |Y_i|$, respectively. Furthermore, by Lemma 3.16 (iii) and (ii),

$$|\hat{\theta}_i| \le \int_{1/n}^1 |\hat{\theta}_i(\tau)| \pi(\tau \mid Y^n)\, d\tau \lesssim t_n |Y_i| e^{Y_i^2/2} + |Y_i| \Pi(\tau \ge 5t_n \mid Y^n) \lesssim t_n |Y_i| e^{Y_i^2/2},$$

since $\Pi(\tau \ge 5t_n \mid Y^n) \lesssim e^{-c_0 p_n} \ll t_n$ if $p_n \gtrsim \log n$.

If the weaker Condition 7 is substituted for Condition 6, then in the preceding we must replace t_n by $(\log n) t_n$. The arguments go through, but with an additional $\log n$ factor in the upper bound on the radius $\hat{r}_i(\alpha)$. This is compensated by the stronger assumption $f_n \gg \log n$ on the lower bound of M_a. □

Technical Lemmas

The next lemma extends Lemma 3.8 to nondeterministic values of τ.

Lemma 3.19. If $n\tau/\zeta_\tau \to \infty$, then for every constant $C > 0$ there exists a constant $D > 0$ such that

$$P_{\theta_0}\Big(\inf_{t\in[C^{-1}\tau, C\tau]} \hat{r}(\alpha, t) \geq D\sqrt{n\tau\zeta_\tau}\Big) \to 1.$$

Proof. Set $T = [C^{-1}\tau, C\tau]$. By the arguments in the proof of Lemma 3.8 and with the same notation, for $1/c^2 \leq 1 - \alpha$,

$$\inf_{t\in T} \hat{r}^2(\alpha, t) \geq \inf_{t\in T} \mathbb{E}(W \mid Y^n, t) - c \sup_{t\in T} \mathrm{sd}(W \mid Y^n, t).$$

By the first assertion of Lemma 3.17 we have $\inf_{t\in T} \mathbb{E}_0\mathbb{E}(W \mid Y^n, t) \gtrsim n\tau\zeta_\tau$. Combination with Lemma 3.34 gives that the infimum on the right side of the display is bounded below by a multiple of $n\tau\zeta_\tau$, with probability tending to one. By the second assertion of Lemma 3.17 we have $\mathbb{E}_0 \sup_{t\in T} \mathrm{var}(W \mid Y^n, t) \lesssim n\tau\zeta_\tau^3$. An application of Markov's inequality shows that the supremum on the right side of the display is bounded above by $o(n\tau\zeta_\tau)$, with probability tending to one, in view of the assumption that $n\tau/\zeta_\tau \to \infty$. □

Lemma 3.20. Suppose that the density of π_n is bounded away from zero on $[1/n, 1]$. For every sufficiently large constant D there exists $d > 0$ such that $\hat{r}(\alpha) \geq d\sqrt{n\zeta_{\tilde{\tau}_n}\tilde{\tau}_n}$ with P_{θ_0}-probability tending to one, uniformly in θ_0 satisfying the excessive-bias restriction (3.14) with $\tilde{p} \geq D \log n$, where $\tilde{\tau}_n = \tau_n(\tilde{p})$.

Proof. Set $T_n = [C^{-1}\tilde{\tau}_n, C\tilde{\tau}_n]$, for C the constant in Lemma 3.21. Then by the definition of $\hat{r}_n(\alpha)$ and the latter lemma $\int_{\tau\in T_n} \Pi\Big(\|\theta - \hat{\theta}\|_2 \leq r(\alpha) \mid \tau, Y^n\Big)\pi(\tau \mid Y^n\Big)\, d\tau$ is equal to $1 - \alpha + o(1)$. Therefore there exists $\tau = \tau(Y^n) \in T_n$ such that

$$\Pi\Big(\|\theta - \hat{\theta}\|_2 \leq \hat{r}(\alpha) \mid \tau, Y^n\Big) \geq 1 - 2\alpha.$$

Introduce the notation $\tilde{W} = \|\theta - \hat{\theta}\|_2^2$, and denote by $\mathbb{E}(\cdot|Y^n, \tau)$ and $\mathrm{sd}(\cdot|Y^n, \tau)$ the posterior expected value and standard variation for given τ. By an application of Chebyshev's inequality, as in the proofs of Lemmas 3.8 and 3.19, we see that $\hat{r}(\alpha) \geq \mathbb{E}(\tilde{W} \mid \tau, Y^n) - c\,\mathrm{sd}(\tilde{W} \mid \tau, Y^n)$, for a sufficiently small constant $c > 0$. Hence it suffices to show that $\inf_{\tau\in T_n} \mathbb{E}(\tilde{W} \mid \tau, Y^n) \gtrsim n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}$ and $\sup_{\tau\in T_n} \mathrm{sd}(\tilde{W} \mid \tau, Y^n) \ll n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}$, with P_{θ_0}-probability tending to one.

Since $\hat{\theta}(\tau)$ is the mean of θ given (Y^n, τ) and the coordinates θ_i are conditionally independent, for $W = \|\theta - \hat{\theta}(\tau)\|_2^2$,

$$\mathbb{E}(\tilde{W} \mid \tau, Y^n) = \mathbb{E}(W \mid \tau, Y^n) + \|\hat{\theta} - \hat{\theta}(\tau)\|_2^2 \geq \mathbb{E}(W \mid \tau, Y^n),$$

$$\mathrm{var}(\tilde{W} \mid \tau, Y^n) \lesssim \sum_{i=1}^n \mathbb{E}\Big((\theta_i - \hat{\theta}_i(\tau))^4 \mid \tau, Y^n\Big) + \sum_{i=1}^n \Big(\hat{\theta}_i - \hat{\theta}_i(\tau)\Big)^4.$$

The proof of Lemma 3.19 shows that $\inf_{\tau\in T_n} \mathbb{E}(W \mid \tau, Y^n) \gtrsim n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}$, with P_{θ_0}-probability tending to one, and hence the same conclusion holds for $\inf_{\tau\in T_n} \mathbb{E}(\tilde{W} \mid \tau, Y^n)$.

It remains to deal with the variance in the preceding display. By Lemma 3.17 the $\mathbb{E}_0$-expected value of the supremum over $\tau \in T_n$ of the first term on the right is bounded above by $n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}^3$, which shows that this term is suitably bounded in view of Markov's inequality. By Jensen's inequality the second can be bounded as

$$\begin{aligned}\|\hat{\theta}(\tau) - \hat{\theta}\|_4^4 &\leq \int_{1/n}^1 \|\hat{\theta}(\tau) - \hat{\theta}(t)\|_4^4 \pi(t \mid Y^n)\, dt \\ &\leq \sup_{t\in T_n} \|\hat{\theta}(\tau) - \hat{\theta}(t)\|_4^4 + \sup_{t\in[1/n,1]} \|\hat{\theta}(\tau) - \hat{\theta}(t)\|_4^4\, \Pi(t \notin T_n \mid Y^n), \end{aligned} \tag{3.34}$$

where $\|\theta\|_4^4 = \sum_{i=1}^n \theta_i^4$. In view of Lemma 3.16 (i)+(ii),

$$\sup_{\tau_1,\tau_2\in[1/n,1]} \|\hat{\theta}(\tau_1) - \hat{\theta}(\tau_2)\|_4^4 \lesssim \sup_{\tau\in[1/n,1]} \|\hat{\theta}(\tau) - Y^n\|_4^4 \lesssim 8n(\log n)^2.$$

Furthermore $\Pi(\tau \notin T_n \mid Y^n) \leq e^{-c_3\tilde{p}}$ by Lemmas 3.21 and 3.6, for a constant $c_3 > 0$. Hence for $\tilde{p} \geq D\log n$, where $D > c_3^{-1}$, the second term on the right hand side of (3.34) tends to zero.

To bound the first term of (3.34) we first use the triangle inequality to obtain that $\sup_{t\in T_n} \|\hat{\theta}(\tau) - \hat{\theta}(t)\|_4 \leq 2\sup_{t\in T_n} \|\hat{\theta}(t) - \theta_0\|_4$. We next split the sum in $\|\hat{\theta}(t) - \theta_0\|_4^4$ in the terms with $|\theta_{0,i}| > \zeta_{\tilde{\tau}_n}/10$ and the remaining terms.

If $|\theta_{0,i}| > \zeta_{\tilde{\tau}_n}/10$, then we use that $|\hat{\theta}_i(t) - \theta_{0,i}| \leq |\hat{\theta}_i(t) - Y_i| + |Y_i - \theta_{0,i}| \lesssim \zeta_t + |Y_i - \theta_{0,i}|$, so that

$$\mathbb{E}_{\theta_{0,i}} \sup_{t\in T_n} |\theta_{0,i} - \hat{\theta}_i(t)|^4 \lesssim \zeta_{\tilde{\tau}_n}^4 + 1 \lesssim \zeta_{\tilde{\tau}_n}^4.$$

By an analogous argument as in the proof of Theorem 3.13 the number of terms with $|\theta_{0,i}| > \zeta_{\tilde{\tau}_n}/10$ is bounded by a multiple of $\tilde{p}$, so that their total contribution is bounded above by $\tilde{p}\zeta_{\tilde{\tau}_n}^4$.

For the terms with $|\theta_{0,i}| \leq \zeta_{\tilde{\tau}_n}/10$, we first use that $|\hat{\theta}_i(t) - \theta_{0,i}| \leq |\hat{\theta}_i(t)| + |\theta_{0,i}| \leq |Y_i - \theta_{0,i}| + 2|\theta_{0,i}|$, so that

$$\mathbb{E}_{\theta_{0,i}} \sup_{t\in T_n} |\hat{\theta}_i(t) - \theta_{0,i}|^4 \mathbf{1}_{|Y_i-\theta_{0,i}|>\zeta_{\tilde{\tau}_n}} \lesssim \int_{\zeta_{\tilde{\tau}_n}}^\infty y^4\varphi(y)\, dy + \theta_{0,i}^4 \lesssim \tilde{\tau}_n\zeta_{\tilde{\tau}_n}^3 + \theta_{0,i}^4.$$

Second we use that $|\hat{\theta}_i(t) - \theta_{0,i}| \lesssim \tau|Y_i|e^{Y_i^2/2} + |\theta_{0,i}|$, by Lemma 3.16 (iii), so that

$$\begin{aligned}\mathbb{E}_{\theta_{0,i}} \sup_{t\in T_n} |\hat{\theta}_i(t) - \theta_{0,i}|^4 \mathbf{1}_{|Y_i-\theta_{0,i}|\leq\zeta_{\tilde{\tau}_n}} &\lesssim \tilde{\tau}_n^4 \int_{-\zeta_{\tilde{\tau}_n}}^{\zeta_{\tilde{\tau}_n}} (y + \theta_{0,i})^4 e^{2(y+\theta_{0,i})^2}\varphi(y)\, dy + \theta_{0,i}^4 \\ &\lesssim \tilde{\tau}_n\zeta_{\tilde{\tau}_n}^3 e^{4\zeta_{\tilde{\tau}_n}|\theta_{0,i}|+2\theta_{0,i}^2} + \theta_{0,i}^4.\end{aligned}$$

For $|\theta_{0,i}| \lesssim \zeta_{\tilde{\tau}_n}^{-1}$, the exponential in the first term is bounded, and the first term is bounded above by $\tilde{\tau}_n\zeta_{\tilde{\tau}_n}^3$. For $|\theta_{0,i}| \gtrsim \zeta_{\tilde{\tau}_n}^{-1}$, but still $|\theta_{0,i}| \leq \zeta_{\tilde{\tau}_n}/10$, the first term can be seen to be bounded above by $\tilde{\tau}_n\zeta_{\tilde{\tau}_n}^3\tilde{\tau}_n^{-21/25}$, which is bounded by $\theta_{0,i}^4$ in that case.

Combining all the preceding computations, we obtain:

$$\mathbb{E}_{\theta_0} \sup_{t\in T_n} \|\theta_0 - \hat{\theta}(t)\|_4^4 \lesssim \tilde{p}\zeta_{\tilde{\tau}_n}^4 + n\tilde{\tau}_n\zeta_{\tilde{\tau}_n}^3 + \sum_{i:|\theta_{0,i}|<\zeta_{\tilde{\tau}_n}/10} \theta_{0,i}^4.$$

We see that this is of the desired order $n\tilde{\tau}_n\zeta^3_{\tilde{\tau}_n}$ by bounding $\theta^4_{0,i}$ by $\zeta^2_{\tilde{\tau}_n}\theta^2_{0,i}$, and next applying the excessive-bias restriction. □

Lemma 3.21. If θ_0 satisfies the excessive-bias restriction (3.14) with $\tilde{p} \geq D\log n$ for a sufficiently large constant D, and the density of π_n is bounded away from zero on $[1/n, 1]$, then there exist constants $C > 0$ and $c_3 > 0$ such that

$$\Pi(\tau : \tau \leq C^{-1}\tilde{\tau}_n \text{ or } \tau \geq C\tilde{\tau}_n \mid Y^n) \lesssim e^{-c_3\tilde{p}}.$$

Proof. As seen in the proof of Lemma 3.12 the function $\tau \mapsto M_\tau(Y^n)$ is increasing for $\tau \leq c_5\tilde{\tau}_n$. Inspection of the proof (see (3.20)) shows that its derivative is bounded below by $c_6\tilde{p}/\tau$ for τ in the interval $[c\tilde{\tau}_n, 2c\tilde{\tau}_n]$, for $2c < c_5/2$ and suitably chosen c_5. This shows that $M_\tau(Y^n) - M_{c\tilde{\tau}_n}(Y^n) \geq c_8\tilde{p}$ in the interval $[2c\tilde{\tau}_n, 4c\tilde{\tau}_n]$, whence

$$\Pi(\tau : \tau \leq c\tilde{\tau}_n \mid Y^n) \leq \frac{\int_{1/n}^{c\tilde{\tau}_n} e^{M_\tau(Y^n)}\pi(\tau)\,d\tau}{\int_{2c\tilde{\tau}_n}^{4c\tilde{\tau}_n} e^{M_\tau(Y^n)}\pi(\tau)\,d\tau} \lesssim \frac{e^{M_{c\tilde{\tau}_n}(Y^n)}}{e^{M_{c\tilde{\tau}_n}(Y^n)+c_8\tilde{p}}c\tilde{\tau}_n}.$$

This is bounded by $e^{-c_3\tilde{p}}$, by the assumption that $\tilde{p} \gtrsim \log n$.

The same bound on $\Pi(\tau : \tau \geq c\tilde{\tau}_n \mid Y^n)$ can be verified following the same reasoning, now using that $\tau \mapsto M_\tau(Y^n)$ is decreasing for $\tau \geq c_6\tilde{\tau}_n$ with derivative bounded above by $-c_9\tilde{p}/\tau$ on an interval $[c\tilde{\tau}_n/2, c\tilde{\tau}_n]$ for $c/2 > 2c_6$ (see (3.21)). □

3.6.6 Lemmas supporting the MMLE results

For $k \in \{-1/2, 1/2, 3/2\}$ define a function $I_k : \mathbb{R} \to \mathbb{R}$ by

$$I_k(y) := \int_0^1 z^k \frac{1}{\tau^2 + (1-\tau^2)z} e^{y^2 z/2}\,dz. \tag{3.35}$$

The Bayesian marginal density of Y_i given τ is the convolution $\psi_\tau := \varphi * g_\tau$ of the standard normal density and the prior density of g_τ, given in (3.3). The latter is a half-Cauchy mixture of normal densities $\varphi_{\tau\lambda}$ with mean zero and standard deviation $\tau\lambda$. By Fubini's theorem it follows that ψ_τ is a half-Cauchy mixture of the densities $\varphi * \varphi_{\tau\lambda}$. In other words

$$\begin{aligned}\psi_\tau(y) &= \int_0^\infty \frac{e^{-\frac{1}{2}y^2/(1+\tau^2\lambda^2)}}{\sqrt{1+\tau^2\lambda^2}\sqrt{2\pi}} \frac{2}{1+\lambda^2}\frac{1}{\pi}\,d\lambda = \int_0^1 \frac{e^{-\frac{1}{2}y^2(1-z)}}{\sqrt{2\pi}\pi} \frac{\tau z^{-1/2}}{\tau^2(1-z)+z}\,dz \\ &= \frac{\tau}{\pi} I_{-1/2}(y)\varphi(y),\end{aligned} \tag{3.36}$$

where the second step follows by the substitution $1 - z = (1+\tau^2\lambda^2)^{-1}$ and some algebra. Note that $I_{-1/2}$ depends on τ, but this has been suppressed from the notation I_k.

Set

$$m_\tau(y) = y^2\frac{I_{1/2}(y) - I_{3/2}(y)}{I_{-1/2}(y)} - \frac{I_{1/2}(y)}{I_{-1/2}(y)}. \tag{3.37}$$

Lemma 3.22. The derivative of the log-likelihood function takes the form

$$\frac{d}{d\tau} M_\tau(y^n) = \frac{1}{\tau} \sum_{j=1}^{n} m_\tau(y_j).$$

Proof. From (3.36) we infer that, with a dot denoting the partial derivative with respect to τ,

$$\frac{\dot\psi_\tau}{\psi_\tau} = \frac{1}{\tau} + \frac{\dot I_{-1/2}}{I_{-1/2}} = \frac{I_{-1/2} + \tau \dot I_{-1/2}}{\tau I_{-1/2}} = \frac{\int_0^1 \frac{e^{y^2 z/2}}{\sqrt{z} N(z)^2} [N(z) - 2\tau^2(1-z)]\, dz}{\tau I_{-1/2}},$$

where $N(z) = \tau^2(1-z) + z = \tau^2 + (1-\tau^2)z$. By integration by parts,

$$y^2 (I_{1/2} - I_{3/2})(y) = \int_0^1 \frac{\sqrt{z}(1-z)}{N(z)} y^2 e^{y^2 z/2}\, dz = -2 \int_0^1 e^{y^2 z/2}\, d\Big[\frac{\sqrt{z}(1-z)}{N(z)} \Big].$$

Substituting the right hand side in formula (3.37), we readily see by some algebra that τ^{-1} times the latter formula reduces to the right side of the preceding display. □

Proposition 3.23. Let $Y \sim N(\theta, 1)$. Then $\sup_{\tau \in [\varepsilon, 1]} \mathbb{E}_0 m_\tau(Y) < 0$ for every $\varepsilon > 0$, and as $\tau \to 0$,

$$\mathbb{E}_\theta m_\tau(Y) = \begin{cases} -\frac{2^{3/2}}{\pi^{3/2}} \frac{\tau}{\zeta_\tau} \big(1 + o(1)\big), & |\theta| = o(\zeta_\tau^{-2}), \\ o(\tau^{1/16} \zeta_\tau^{-1}), & |\theta| \le \zeta_\tau / 4. \end{cases} \tag{3.38}$$

Proof. Let κ_τ be the solution to the equation $e^{y^2/2}/(y^2/2) = 1/\tau$, that is

$$e^{\kappa_\tau^2/2} = \frac{1}{\tau} \kappa_\tau^2/2, \qquad \kappa_\tau \sim \zeta_\tau + \frac{2 \log \zeta_\tau}{\zeta_\tau}, \qquad \zeta_\tau = \sqrt{2 \log(1/\tau)}.$$

We split the integral over $(0, \infty)$ into the three parts $(0, \zeta_\tau)$, $(\zeta_\tau, \kappa_\tau)$, and (κ_τ, ∞), where we shall see that the last two parts give negligible contributions.

By Lemma 3.28(vi) and (vii), if $|\theta| \kappa_\tau = O(1)$,

$$\int_{|y| \ge \kappa_\tau} m_\tau(y) \varphi(y - \theta)\, dy \lesssim \int_{z \ge \kappa_\tau - |\theta|} \varphi(z)\, dz \lesssim \frac{e^{-(\kappa_\tau - \theta)^2/2}}{\kappa_\tau - \theta} \lesssim \frac{e^{-\kappa_\tau^2/2}}{\kappa_\tau},$$

$$\int_{\zeta_\tau \le |y| \le \kappa_\tau} m_\tau(y) \varphi(y - \theta)\, dy \lesssim \int_{\zeta_\tau \le |y| \le \kappa_\tau} \frac{\tau e^{y^2/2 - (y-\theta)^2/2}}{y^2}\, dy \lesssim \frac{\tau(\kappa_\tau - \zeta_\tau)}{\zeta_\tau^2}.$$

By the definition of κ_τ, both terms are of smaller order than τ/ζ_τ.

Because $e^{y^2/2}/y^2$ is increasing for large y and reaches the value τ^{-1}/ζ_τ^2 at $y = \zeta_\tau$, Lemma 3.30 gives that $I_{-1/2}(y) = \pi \tau^{-1}(1 + O(1/\zeta_\tau^2))$ uniformly in y in the interval $(0, \zeta_\tau)$. Therefore

$$\int_{|y| \le \zeta_\tau} m_\tau(y) \varphi(y - \theta)\, dy = \int_0^{\zeta_\tau} \frac{y^2 I_{1/2}(y) - y^2 I_{3/2}(y) - I_{1/2}(y)}{\tau^{-1} \pi} \varphi(y)\, dy + R_\tau,$$

where the remainder R_τ is bounded in absolute value by $\int_0^{\zeta_\tau} |y^2(I_{1/2} - I_{3/2})(y) - I_{1/2}(y)|$ $\varphi(y)\,dy$ times $\sup_{0\le y\le\zeta_\tau}\Big|\varphi(y-\theta)/(I_{-1/2}(y)\varphi(y)) - 1/(\tau^{-1}\pi)\Big|$, which is bounded above by $\tau\big(\zeta_\tau^{-2} + e^{|\theta|\zeta_\tau - \theta^2/2} - 1\big) = o(\tau\zeta_\tau^{-1})$, for $|\theta| = o(\zeta_\tau^{-2})$. By Lemma 3.31 the integrand in the integral is bounded above by a constant for y near 0 and by a multiple of y^{-2} otherwise, and hence the integral remains bounded. Thus the remainder R_τ is negligible. By Fubini's theorem the integral in the preceding display can be rewritten

$$\begin{aligned}\frac{\tau}{\pi}\int_0^1 \frac{\sqrt{z}}{\tau^2 + (1-\tau^2)z}\int_0^{\zeta_\tau}\Big[y^2(1-z) - 1\Big]\frac{e^{-y^2(1-z)/2}}{\sqrt{2\pi}}\,dy\,dz \\ = -\frac{\tau}{\pi}\int_0^1 \frac{\sqrt{z}}{\tau^2 + (1-\tau^2)z}\int_{\zeta_\tau}^{\infty}\Big[y^2(1-z) - 1\Big]\frac{e^{-y^2(1-z)/2}}{\sqrt{2\pi}}\,dy\,dz\end{aligned}$$

by the fact that the inner integral vanishes when computed over the interval $(0,\infty)$ rather than $(0,\zeta_\tau)$. Since $\int_y^\infty [(va)^2 - 1]\varphi(va)\,dv = y\varphi(ya)$, it follows that the right side is equal to

$$-\frac{\tau}{\pi}\int_0^1 \frac{\sqrt{z}}{\tau^2 + (1-\tau^2)z}\frac{\zeta_\tau\, e^{-\zeta_\tau^2(1-z)/2}}{\sqrt{2\pi}}\,dz.$$

We split the integral in the ranges $(0,1/2)$ and $(1/2,1)$. For z in the first range we have $1 - z \ge 1/2$, whence the contribution of this range is bounded in absolute value by

$$\frac{\zeta_\tau\tau}{\pi\sqrt{2\pi}}e^{-\zeta_\tau^2/4}\int_0^{1/2}\frac{\sqrt{z}}{(1-\tau^2)z}\,dz = O(\zeta_\tau\tau e^{-\zeta_\tau^2/4}).$$

Uniformly in z in the range $(1/2,1)$ we have $\tau^2 + (1-\tau^2)z \sim z$, and the corresponding contribution is

$$-\frac{\tau}{\pi}\int_{1/2}^1 \frac{1}{\sqrt{z}}\frac{\zeta_\tau\, e^{-\zeta_\tau^2(1-z)/2}}{\sqrt{2\pi}}\,dz = -\frac{\tau}{\pi\zeta_\tau\sqrt{2\pi}}\int_0^{\zeta_\tau^2/2}\frac{1}{\sqrt{1 - u/\zeta_\tau^2}}e^{-u/2}\,du.$$

by the substitution $\zeta_\tau^2(1-z) = u$. The integral tends to $\int_0^\infty e^{-u/2}\,du = 2$, and hence the expression is asymptotic to half the expression as claimed.

The second statement follows by the same estimates, where now we use that $e^{|\theta|2\zeta_\tau - \theta^2/2} \le \tau^{-15/16}$, if $|\theta| \le \zeta_\tau/4$.

Since $\mathbb{E}_0 m_\tau(Y) \sim -c\tau/\zeta_\tau$ for a positive constant c, as $\tau \downarrow 0$, the continuous function $\tau \mapsto \mathbb{E}_0 m_\tau(Y)$ is certainly negative if $\tau > 0$ and τ is close to zero. To see that it is bounded away from zero as τ moves away from 0, we computed $\mathbb{E}_0 m_\tau(Y)$ via numerical integration. The result is shown in Figure 3.7. □

Lemma 3.24. For any $\varepsilon_\tau \downarrow 0$ and uniformly in $I_0 \subseteq \{i : |\theta_{0,i}| \le \zeta_\tau^{-1}\}$ with $|I_0| \gtrsim n$,

$$\sup_{1/n\le\tau\le\varepsilon_\tau}\frac{1}{|I_0|}\Big|\sum_{i\in I_0} m_\tau(Y_i)\frac{\zeta_\tau}{\tau} - \sum_{i\in I_0}\mathbb{E}_{\theta_0} m_\tau(Y_i)\frac{\zeta_\tau}{\tau}\Big| \xrightarrow{P_{\theta_0}} 0.$$

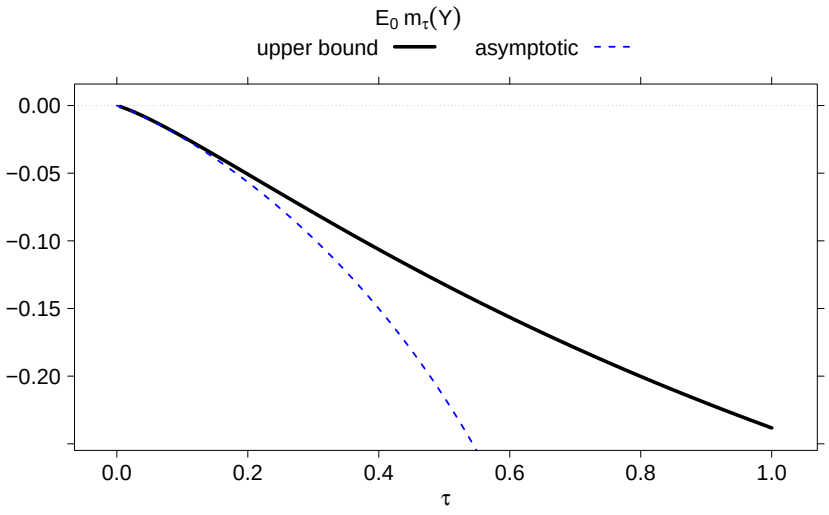

Figure 3.7: Upper bound on $\mathbb{E}_0 m_\tau(Y)$ as computed with the R `integrate()` routine (solid line). The upper bound $m_\tau(y) \le y^2$ was used for $|y| > 500$ for numerical stability. The dashed line shows the asymptotic value (3.38).

Similarly, uniformly in $I_1 \subseteq \{i : |\theta_{0,i}| \le \zeta_\tau/4\}$,

$$\sup_{1/n \le \tau \le \varepsilon_\tau} \frac{1}{|I_1|} \Big| \sum_{i \in I_1} m_\tau(Y_i) \frac{\zeta_\tau}{\tau^{1/32}} - \sum_{i \in I_1} \mathbb{E}_{\theta_0} m_\tau(Y_i) \frac{\zeta_\tau}{\tau^{1/32}} \Big| \overset{P_{\theta_0}}{\to} 0.$$

Proof. Write $G_n(\tau) = |I_0|^{-1} \sum_{i \in I_0} m_\tau(Y_i)(\zeta_\tau)/\tau$. In view of Corollary 2.2.5 of Van der Vaart and Wellner (1996) (applied with $\psi(x) = x^2$) it is sufficient to show that $\operatorname{var}_{\theta_0} G_n(\tau) \to 0$ for some τ, and

$$\int_0^{\operatorname{diam}_n} \sqrt{N(\varepsilon, [1/n, 1], d_n)}\, d\varepsilon = o(1), \tag{3.39}$$

where d_n is the intrinsic metric defined by its square $d_n^2(\tau_1, \tau_2) = \operatorname{var}_{\theta_0}\big(G_n(\tau_1) - G_n(\tau_2)\big)$, diam_n is the diameter of the interval $[1/n, 1]$ with respect to the metric d_n, and $N(\varepsilon, A, d_n)$ is the covering number of the set A with ε radius balls with respect to the metric d_n.

If $|\theta_{0,i}| \le \zeta_\tau^{-1}$, then in view of Lemma 3.26, as $\tau \to 0$,

$$\operatorname{var}_{\theta_0} G_n(\tau) \le \frac{1}{|I_0|} \mathbb{E}_{\theta_0}\big(m_\tau(Y)\zeta_\tau/\tau\big)^2 = o(\tau^{-1}/|I_0|).$$

This tends to zero, as $\tau n \ge 1$ by assumption. Combining this with the triangle inequality we also see that the diameter diam_n tends to 0.

Next we deal with the entropy. The metric d_n is up to a constant equal to the square root of the left side of (3.40). By Lemma 3.25 it satisfies

$$d_n(\tau_1, \tau_2) \lesssim |I_0|^{-1/2} |\tau_2/\tau_1 - 1| \tau_1^{-1/2}.$$

To compute the covering number of the interval $[1/n,1]$, we cover this by dyadic blocks $[2^i/n, 2^{i+1}/n]$, for $i = 0,1,2,\ldots,\log_2 n$. On the ith block the distance $d_n(\tau_1,\tau_2)$ is bounded above by a multiple of $n|\tau_1-\tau_2|/2^{3i/2}$. We conclude that the ith block can be covered by a multiple of $\varepsilon^{-1}2^{-i/2}$ balls of radius ε. Therefore the whole interval $[1/n,1]$ can be covered by a multiple of $\varepsilon^{-1}\sum_i 2^{-i/2} \lesssim \varepsilon^{-1}$ balls of radius ε. Hence the integral of the entropy is bounded by

$$\int_0^{\mathrm{diam}_n} \sqrt{N(\varepsilon,[1/n,1],d_n)}\,d\varepsilon \lesssim \int_0^{\mathrm{diam}_n} \varepsilon^{-1/2}\,d\varepsilon.$$

This tends to zero as diam_n tends to zero.

The second assertion of the lemma follows similarly, where we use the second parts of Lemmas 3.26 and 3.25. □

Lemma 3.25. Let $Y \sim N(\theta,1)$. For $|\theta| \lesssim \zeta_\tau^{-1}$ and $0 < \tau_1 < \tau_2 \le 1/2$,

$$\mathbb{E}_\theta\left(\frac{\zeta_{\tau_1}}{\tau_1}m_{\tau_1}(Y) - \frac{\zeta_{\tau_2}}{\tau_2}m_{\tau_2}(Y)\right)^2 \lesssim (\tau_2-\tau_1)^2\tau_1^{-3}. \tag{3.40}$$

Furthermore, for $|\theta| \le \zeta_\tau/4$, and $\varepsilon = 1/16$ and $0 < \tau_1 < \tau_2 \le 1/2$,

$$\mathbb{E}_\theta\left(\frac{\zeta_{\tau_1}}{\tau_1^\varepsilon}m_{\tau_1}(Y) - \frac{\zeta_{\tau_2}}{\tau_2^\varepsilon}m_{\tau_2}(Y)\right)^2 \lesssim (\tau_2-\tau_1)^2\tau_1^{-2-\varepsilon}.$$

Proof. In view of Lemma 3.33 the left side of (3.40) is bounded above by, for $\dot m_\tau$ denoting the partial derivative of m_τ with respect to τ,

$$\begin{aligned}
&(\tau_1-\tau_2)^2 \sup_{\tau\in[\tau_1,\tau_2]} \mathbb{E}_\theta\Big(\frac{\zeta_\tau}{\tau}\dot m_\tau(Y) - \frac{\zeta_\tau+\zeta_\tau^{-1}}{\tau^2}m_\tau(Y)\Big)^2 \\
&\quad\le (\tau_1-\tau_2)^2\Big[2\sup_{\tau\in[\tau_1,\tau_2]} \mathbb{E}_\theta\Big(\frac{\zeta_\tau}{\tau}\dot m_\tau(Y)\Big)^2 + 2\sup_{\tau\in[\tau_1,\tau_2]} \mathbb{E}_\theta\Big(\frac{\zeta_\tau+\zeta_\tau^{-1}}{\tau^2}m_\tau(Y)\Big)^2\Big].
\end{aligned}$$

By Lemma 3.26 the second expected value on the right hand side is bounded from above by a multiple of $\sup_{\tau\in[\tau_1,\tau_2]}\tau^{-3} \lesssim \tau_1^{-3}$.

To handle the first expected value, we note that the partial derivative of I_k with respect to τ is given by $\dot I_k = 2\tau(J_{k+1}-J_k)$, for

$$J_k(y) = \int_0^1 \frac{z^k}{(\tau^2+(1-\tau^2)z)^2}e^{y^2z/2}dz. \tag{3.41}$$

Therefore, by (3.37),

$$\begin{aligned}
\dot m_\tau(y) &= (y^2-1)\frac{\dot I_{1/2}}{I_{-1/2}}(y) - y^2\frac{\dot I_{3/2}}{I_{-1/2}}(y) - \frac{\dot I_{1/2}}{I_{-1/2}}(y)m_\tau(y) \\
&= 2\tau\Big[(y^2-1)\frac{J_{3/2}-J_{1/2}}{I_{-1/2}}(y) - y^2\frac{J_{5/2}-J_{3/2}}{I_{-1/2}}(y) - \frac{J_{1/2}-J_{-1/2}}{I_{-1/2}}(y)m_\tau(y)\Big].
\end{aligned}$$

Since $J_k \le I_{k-1}/(1-\tau^2)$ and $J_k \le I_k/\tau^2$, and $k \mapsto I_k$ and $k \mapsto J_k$ are decreasing and nonnegative, we have that

$$0 \le \frac{J_{3/2} - J_{5/2}}{I_{-1/2}} \le \frac{J_{1/2} - J_{3/2}}{I_{-1/2}} \le \frac{J_{1/2}}{I_{-1/2}} \le 4,$$
$$0 \le \frac{J_{-1/2} - J_{1/2}}{I_{-1/2}} \le \frac{J_{-1/2}}{I_{-1/2}} \le \frac{1}{\tau^2}. \tag{3.42}$$

By combining the preceding two displays we conclude

$$\mathbb{E}_\theta \dot{m}_\tau^2(Y) \lesssim \tau^2 \Big[1 + \mathbb{E}_\theta Y^4 + \frac{1}{\tau^4} \mathbb{E}_\theta m_\tau^2(Y) \Big]. \tag{3.43}$$

Here $\mathbb{E}_\theta Y^4$ is bounded and $\mathbb{E}_\theta m_\tau^2(Y)$ is bounded above by $\tau\zeta_\tau^{-2}$ by Lemma 3.26. It follows that $(\zeta_\tau/\tau)^2 \mathbb{E}_\theta \dot{m}_\tau^2(Y)$ is bounded by a multiple of $\tau^{-3} \le \tau_1^{-3}$.

For the proof of the second assertion of the lemma, when $|\theta| \le \zeta_\tau/4$, we argue similarly, but now must bound,

$$(\tau_1 - \tau_2)^2 \Big[2 \sup_{\tau \in [\tau_1, \tau_2]} \mathbb{E}_\theta \Big(\frac{\zeta_\tau}{\tau^\varepsilon} \dot{m}_\tau(Y) \Big)^2 + 2 \sup_{\tau \in [\tau_1, \tau_2]} \mathbb{E}_\theta \Big(\frac{\varepsilon\zeta_\tau + \zeta_\tau^{-1}}{\tau^{1+\varepsilon}} m_\tau(Y) \Big)^2 \Big].$$

The same arguments as before apply, now using the second bound from Lemma 3.26. □

Lemma 3.26. Let $Y \sim N(\theta, 1)$. Then, as $\tau \to 0$,

$$\mathbb{E}_\theta m_\tau^2(Y) = \begin{cases} o(\tau\zeta_\tau^{-2}), & |\theta| \lesssim \zeta_\tau^{-1}, \\ o(\tau^{1/16}\zeta_\tau^{-2}), & |\theta| \le \zeta_\tau/4. \end{cases}$$

Proof. By Lemma 3.28 (i), (vi) and (vii) we have, if $|\theta|\zeta_\tau \lesssim 1$,

$$\int_{|y| \ge \kappa_\tau} m_\tau^2(y) \varphi(y-\theta)\, dy \lesssim \int_{|z| \ge \kappa_\tau - \theta}^{\infty} \varphi(z)\, dz \lesssim e^{-(\kappa_\tau - \theta)^2/2} (\kappa_\tau - \theta)^{-1} \lesssim \tau\zeta_\tau^{-3},$$
$$\int_{\zeta_\tau \le |y| \le \kappa_\tau} m_\tau^2(y) \varphi(y-\theta)\, dy \lesssim \int_{\zeta_\tau}^{\kappa_\tau} \tau y^{-2} e^{y^2/2 - (y-\theta)^2/2}\, dy = \tau(\kappa_\tau - \zeta_\tau)\zeta_\tau^{-2},$$
$$\int_{|y| \le \zeta_\tau} m_\tau^2(y) \varphi(y-\theta)\, dy \lesssim \tau^2 \int_0^{\zeta_\tau} (y^{-4} \wedge 1) e^{y^2/2} e^{\theta\zeta_\tau - \theta^2/2}\, dy \lesssim \tau\zeta_\tau^{-4}.$$

All three expressions on the right are $o(\tau\zeta_\tau^{-2})$.

The second assertion of the lemma follows by the same inequalities, together with the inequalities $e^{-(\kappa_\tau - \theta)^2/2} \le \tau^{-9/32}$ and $e^{|\theta|2\zeta_\tau - \theta^2/2} \le \tau^{-15/16}$, if $|\theta| \le \zeta_\tau/4$. □

Lemma 3.27. If the cardinality of $I_0 := \{i : \theta_{0,i} = 0\}$ tends to infinity, then

$$\sup_{1/n \le \tau \le 1} \frac{1}{|I_0|} \Big| \sum_{i \in I_0} m_\tau(Y_i) - \sum_{i \in I_0} \mathbb{E}_{\theta_0} m_\tau(Y_i) \Big| \overset{P_{\theta_0}}{\to} 0.$$

Proof. By Lemma 3.28(i) we have that $\mathbb{E}_0 m_\tau^2(Y_i) \lesssim 1$ uniformly in τ and by the proof of Lemma 3.25 $\mathbb{E}_0(m_{\tau_1} - m_{\tau_2})^2(Y_i) \lesssim |\tau_1 - \tau_2|^2/\tau_1$, uniformly in $0 < \tau_1 < \tau_2 \leq 1$. The first shows that the marginal variances of the process $G_n(\tau) := |I_0|^{-1} \sum_{i \in I_0} m_\tau(Y_i)$ tend to zero as $|I_0| \to \infty$. The second allows to control the entropy integral of the process and complete the proof, in the same way as the proof of Lemma 3.24. □

Lemma 3.28. The function $y \mapsto m_\tau(y)$ is symmetric about 0 and nondecreasing on $[0, \infty)$ with

(i) $-1 \leq m_\tau(y) \leq C_u$, for all $y \in \mathbb{R}$ and all $\tau \in [0,1]$, and some $C_u < \infty$.

(ii) $m_\tau(0) = -(2\tau/\pi)(1 + o(1))$, as $\tau \to 0$.

(iii) $m_\tau(\zeta_\tau) = 2/(\pi\zeta_\tau^2)(1 + o(1))$, as $\tau \to 0$.

(iv) $m_\tau(\kappa_\tau) = 1/(\pi + 1)/(1 + o(1))$, as $\tau \to 0$.

(v) $\sup_{y \geq A\zeta_\tau} |m_\tau(y) - 1| = O(\zeta_\tau^{-2})$, as $\tau \to 0$, for every $A > 1$.

(vi) $m_\tau(y) \sim \tau e^{y^2/2}/(\pi y^2/2 + \tau e^{y^2/2})$, as $\tau \to 0$, uniformly in $|y| \geq 1/\varepsilon_\tau$, for any $\varepsilon_\tau \downarrow 0$.

(vii) $|m_\tau(y)| \lesssim \tau e^{y^2/2}(y^{-2} \wedge 1)$, as $\tau \to 0$, for every y.

Proof. As seen in the proof of Lemma 3.22 the function m_τ can be written

$$m_\tau(y) = 1 + \tau \frac{\dot{I}_{-1/2}}{I_{-1/2}}(y) = 1 + 2\tau^2 \int_0^1 \frac{z-1}{\tau^2 + (1-\tau^2)z} g_y(z)\, dz,$$

for $z \mapsto g_y(z)$ the probability density function on $[0,1]$ with $g_y(z) \propto e^{y^2/2} z^{-1/2}/(\tau^2 + (1 - \tau^2)z)$. If y increases, then the probability distribution increases stochastically, and hence so does the expectation of the increasing function $z \mapsto (z-1)/(\tau^2 + (1-\tau^2)z)$. (More precisely, note that g_{y_2}/g_{y_1} is increasing if $y_2 > y_1$ and apply Lemma 3.18.)

(i). The inequality $m_\tau(y) \geq -1$ is immediate from the definition of (3.37) of m_τ and the fact that $I_{3/2} \leq I_{1/2} \leq I_{-1/2}$. For the upper bound it suffices to show that both $\sup_y m_\tau(y)$ remains bounded as $\tau \to 0$ and that $\sup_y \sup_{\tau \geq \delta} m_\tau(y) < \infty$ for every $\delta > 0$.

The first follows from the monotonicity and (v).

For the proof of the second we note that if $\tau \geq \delta > 0$, then $\delta^2 \leq \tau^2 + (1-\tau^2)z \leq 1$, for every $z \in [0,1]$, so that the denominators in the integrands of $I_{-1/2}, I_{1/2}, I_{3/2}$ are uniformly bounded away from zero and infinity and hence

$$m_\tau(y) \leq y^2 \frac{I_{1/2}(y) - I_{3/2}(y)}{I_{-1/2}(y)} \leq \frac{1}{\delta^2} \frac{y^2 \int_0^1 \sqrt{z}(1-z) e^{y^2 z/2}\, dz}{\int_0^1 z^{-1/2} e^{y^2 z/2}\, dz}.$$

After changing variables $zy^2/2 = v$, the numerator and denominator take the forms of the integrals in the second and first assertions of Lemma 3.29, except that the range of integration is $(0, y^2/2)$ rather than $(1, y)$. In view of the lemma the quotient approaches 1 as $y \to \infty$. For y in a bounded interval the leading factor y^2 is bounded, while the integral

in the numerator is smaller than the integral in the denominator, as $z(1-z) \le z \le z^{-1/2}$, for $z \in [0,1]$.

Assertions (ii)-(v) are consequences of the representation (3.37), Lemmas 3.30 and 3.31 and the fact that $I_{1/2}(0) = \int_0^1 z^{-1/2} dz\Big(1 + O(\tau^2)\Big) \to 2$.

Assertions (vi) and (vii) are immediate from Lemmas 3.30 and 3.31. □

Technical lemmas

Lemma 3.29. For any k, as $y \to \infty$,

$$\int_1^y u^k e^u \, du = y^k e^y \Big(1 - k/y + O(1/y^2)\Big).$$

Consequently, as $y \to \infty$,

$$\int_1^y u^k e^u \, du - \frac{1}{y} \int_1^y u^{k+1} e^u \, du = y^{k-1} e^y \Big(1 + O(1/y)\Big).$$

Proof. By integrating by parts twice, the first integral is seen to be equal to

$$y^k e^y - e - k y^{k-1} e^y + ke + R,$$

where R satisfies

$$\begin{aligned} |R| &= |k(k-1)| \int_1^y u^{k-2} e^u \, du \\ &\le |k(k-1)| \int_1^{y/2} (1 \vee (y/2)^{k-2}) e^u \, du + |k(k-1)| \int_{y/2}^y ((y/2)^{k-2} \vee y^{k-2}) e^u \, du \\ &\lesssim |k(k-1)| \Big[(1 \vee y^{k-2}) e^{y/2} + y^{k-2} e^y \Big]. \end{aligned}$$

The second assertion follows by applying the first one twice. □

Lemma 3.30. There exist functions R_τ with $\sup_y |R_\tau(y)| = O(\sqrt{\tau})$ as $\tau \downarrow 0$, such that

$$I_{-1/2}(y) = \left(\frac{\pi}{\tau} + \sqrt{y^2/2} \int_1^{y^2/2} \frac{1}{v^{3/2}} e^v \, dv \right) \Big(1 + R_\tau(y)\Big).$$

Furthermore, given $\varepsilon_\tau \to 0$ there exist functions S_τ with $\sup_{y \ge 1/\varepsilon_\tau} |S_\tau(y)| = O(\sqrt{\tau} + \varepsilon_\tau^2)$, such that, as $\tau \downarrow 0$,

$$I_{-1/2}(y) = \left(\frac{\pi}{\tau} + \frac{e^{y^2/2}}{y^2/2} \right) \Big(1 + S_\tau(y)\Big).$$

Proof. For the proof of the first assertion we separately consider the ranges $|y| \le 2\zeta_\tau$ and $|y| > 2\zeta_\tau$. For $|y| \le 2\zeta_\tau$ we split the integral in the definition of $I_{-1/2}$ over the intervals $(0, \tau)$, $(\tau, (2/y^2) \wedge 1)$ and $((2/y^2) \wedge 1, 1)$, where we consider the third interval empty if

$y^2/2 \le 1$. Making the changes of coordinates $z = u\tau^2$ in the first integral, and $(y^2/2)z = v$ in the second and third integrals, we see that

$$I_{-1/2}(y) = \frac{1}{\tau}\int_0^{1/\tau} \frac{1}{\sqrt{u}}\frac{1}{1+(1-\tau^2)u} e^{y^2\tau^2 u/2}\,du + \sqrt{y^2/2}\Big[\int_{y^2\tau/2}^{y^2/2\wedge 1} + \int_{y^2/2\wedge 1}^{y^2/2}\Big]\frac{1}{\sqrt{v}}\frac{1}{\tau^2 y^2/2 + (1-\tau^2)v} e^v\,dv$$

For $|y| \le 2\zeta_\tau$, the exponential in the first integral tends to 1, uniformly in $u \le 1/\tau$. Since $e^u - 1 \le ue^u$, for $u \ge 0$, replacing it by 1 gives an error of at most

$$\frac{1}{\tau}\int_0^{1/\tau} \frac{1}{\sqrt{u}}\frac{e^{y^2\tau/2}y^2\tau^2 u}{1+(1-\tau^2)u}\,du \lesssim \frac{1}{\tau}y^2\tau^{3/2}.$$

As $(1-\tau^2)(1+u) \le 1 + (1-\tau^2)u \le 1+u$, dropping the factor $1-\tau^2$ from the denominator makes a multiplicative error of order $1 + O(\tau^2)$. Since $\int_0^\infty u^{-1/2}/(1+u)\,du = \pi$ and $\int_{1/\tau}^\infty u^{-1/2}/(1+u)\,du \lesssim \tau^{1/2}$, the first term gives a contribution of $\pi/\tau + O(\tau^{-1/2})$, uniformly in $|y| \le 2\zeta_\tau$. In the second integral we bound the factor $\tau^2y^2/2 + (1-\tau^2)v$ below by $(1-\tau^2)v$, the exponential e^v above by e and the upper limit of the integral by 1, and next evaluate the integral to be bounded by a constant times $\tau^{-1/2}$. For the third integral we separately consider the cases that $y^2/2 \le 1$ and $y^2/2 > 1$. In the first case the third integral contributes nothing; the second term (the integral) in the assertion of the lemma is bounded and hence also contributes a negligible amount relative to π/τ. Finally consider the case that $y^2/2 > 1$. If in the third integral we replace $\tau^2y^2/2 + (1-\tau^2)v$ by v, we obtain the second term in the assertion of the lemma. The difference is bounded above by

$$\sqrt{y^2/2}\int_1^{y^2/2}\frac{1}{\sqrt{v}}\frac{\tau^2v + \tau^2y^2}{v(\tau^2y^2/2 + (1-\tau^2)v)}e^v\,dv \lesssim \tau^2\sqrt{y^2/2}\int_1^{y^2/2}(v^{-3/2} + y^2v^{-5/2})e^v\,dv.$$

This is negligible relative to the integral in the assertion. This concludes the proof of the first assertion of the lemma for the range $|y| \le 2\zeta_\tau$.

For $|y|$ in the interval $(2\zeta_\tau, \infty)$ we split the integral in the definition of $I_{-1/2}$ into the ranges $[0, 1/3]$ and $(1/3, 1]$. The contribution of the first range is bounded above by

$$\frac{1}{\tau^2}e^{y^2/6}\int_0^{1/3} z^{-1/2}\,dz \ll \sqrt{\tau}\frac{e^{y^2/2}}{y^2/2},$$

for $|y| \ge 2\zeta_\tau$. This is negligible relative to the integral in the assertion, which expands as $e^{y^2/2}/\sqrt{y^2/2}$, as claimed by the second assertion of the lemma. In the contribution of the second range we use that $z \le \tau^2 + (1-\tau^2)z \le (1+2\tau^2)z$, for $z \ge 1/3$, and see that this is up to a multiplicative term of order $1 + O(\tau^2)$ equal to

$$\int_{1/3}^1 z^{-3/2}e^{y^2z/2}\,dz = \sqrt{y^2/2}\Big[\int_1^{y^2/2} - \int_1^{y^2/6}\Big]v^{-3/2}e^v\,dv.$$

Applying Lemma 3.29, we see that the contribution of the second integral is bounded above by a multiple of $(y^2/2)^{-1}e^{y^2/6}$, which is negligible relative to the first.

To prove the second assertion of the lemma we expand the integral in the first assertion with the help of Lemma 3.29. □

Lemma 3.31. For $k > 0$, there exist functions $R_{\tau,k}$ with $\sup_y |R_{\tau,k}(y)| = O(\tau^{2k/(k+1)})$, and for given $\varepsilon_\tau \to 0$ functions $S_{\tau,k}$ with $\sup_{y \geq 1/\varepsilon_\tau} |S_{\tau,k}(y)| = O(\tau^{2k/(2k+1)} + \varepsilon_\tau^2)$, such that, as $\tau \downarrow 0$,

$$I_k(y) = \frac{1}{(y^2/2)^k} \int_0^{y^2/2} v^{k-1} e^v \, dv \Big(1 + R_{\tau,k}(y)\Big) \lesssim \big(1 \wedge y^{-2}\big) e^{y^2/2},$$
$$I_k(y) = \frac{e^{y^2/2}}{y^2/2} \Big(1 + S_{\tau,k}(y)\Big).$$

There also exist functions $\bar{R}_\tau$ with $\sup_y |\bar{R}_\tau(y)| = O(\tau^{1/2})$ and $\bar{S}_\tau$ with $\sup_{y \geq 1/\varepsilon_\tau} |\bar{S}_\tau(y)| = O(\sqrt{\tau} + \varepsilon_\tau^2)$, such that, as $\tau \downarrow 0$ and $\varepsilon_\tau \to 0$,

$$I_{1/2}(y) - I_{3/2}(y) = \frac{1}{\sqrt{y^2/2}} \int_0^{y^2/2} \frac{1 - 2v/y^2}{\sqrt{v}} e^v \, dv \Big(1 + \bar{R}_\tau(y)\Big) \lesssim (1 \wedge y^{-4}) e^{y^2/2},$$
$$I_{1/2}(y) - I_{3/2}(y) = \frac{e^{y^2/2}}{(y^2/2)^2} \Big(1 + \bar{S}_\tau(y)\Big).$$

Proof. We split the integral in the definition of I_k over the intervals $[0, \tau^a]$ and $[\tau^a, 1]$, for $a = 2/(k+1)$. The contribution of the first integral is bounded above by

$$e^{\tau^a y^2/2} \int_0^{\tau^a} \frac{z^k}{(1-\tau^2)z} \, dz \lesssim e^{\tau^a y^2/2} \tau^{ka}.$$

In the second integral we use that $z \leq \tau^2 + (1-\tau^2)z \leq (\tau^{2-a} + 1 - \tau^2)z$, for $z \geq \tau^a$, to see that the integral is $1 + O(\tau^{2-a})$ times

$$\int_{\tau^a}^1 \frac{z^k}{z} e^{y^2 z/2} \, dz \gtrsim e^{\tau^a y^2/2}.$$

Combining these displays, we see that

$$I_k(y) = \int_{\tau^a}^1 z^{k-1} e^{y^2 z/2} \, dz (1 + O(\tau^{2-a}) + O(\tau^{ka})).$$

This remains valid if we enlarge the range of integration to $[0,1]$. The change of coordinates $zy^2/2 = v$ completes the proof of the equality in the first assertion.

For the second assertion we expand the integral in the first assertion with the help of the second assertion of Lemma 3.29. Note here that for $k > -1$ the integrals in the latter lemma can be taken over $(0, y)$ instead of $(1, y)$, since the difference is a constant.

The inequality in the first assertion is valid for $y \to \infty$, in view of the second assertion, and from the fact that $G(y) := (y^2/2)^{-k} \int_0^{y^2/2} v^{k-1} e^v \, dv$ possesses a finite limit as $y \downarrow 0$ it

follows that it is also valid for $y \to 0$. For intermediate y the inequality follows since the continuous function $y \mapsto G(y)e^{-y^2/2}/(y^{-2} \wedge 1)$ is bounded on compacta in $(0, \infty)$.

For the proofs of the assertions concerning $I_{1/2} - I_{3/2}$ we write

$$I_{1/2}(y) - I_{3/2}(y) = \Big(\int_0^\tau + \int_\tau^1 \Big) \frac{\sqrt{z}(1-z)}{\tau^2 + (1-\tau^2)z} e^{y^2 z/2}\, dz.$$

Next we follow the same approach as previously. □

Lemma 3.32. For any $M \to \infty$ and $\tau_0 > 0$ there exists a constant A such that $0 \leq 1 - I_{1/2}/I_{-1/2}(y) \leq A(\log|y|)/y^2$, for every $|y| \geq M$ and $\tau \geq \tau_0$.

Proof. The first inequality is clear from the fact that $I_{1/2} \leq I_{-1/2}$. For the proof of the upper bound we write the difference $1 - I_{1/2}/I_{-1/2}(y)$ as

$$\frac{\int_0^1 (1-z)z^{-1/2}(\tau^2 + (1-\tau^2)z)^{-1}\, e^{zy^2/2}\, dz}{\int_0^1 z^{-1/2}(\tau^2 + (1-\tau^2)z)^{-1} e^{zy^2/2}\, dz} \leq \frac{\int_0^c (1-z)z^{-1/2}\tau^{-2}\, dz\, e^{cy^2/2}}{\int_d^1 z^{-1/2}\, dz e^{dy^2/2}} + 1 - c.$$

The integral in the numerator is uniformly bounded, while the integral in the denominator is bounded below by a multiple of $1 - d$. We now choose $1 - c = 4\log(y^2/2)/(y^2/2) = 2(1-d)$. □

Lemma 3.33. For any stochastic process $(V_\tau : \tau > 0)$ with continuously differentiable sample paths $\tau \mapsto V_\tau$, with derivative written as $\dot{V}_\tau$,

$$\mathbb{E}(V_{\tau_2} - V_{\tau_1})^2 \leq (\tau_2 - \tau_1)^2 \sup_{\tau \in [\tau_1, \tau_2]} \mathbb{E}\dot{V}_\tau^2.$$

Proof. By the Newton-Leibniz formula, the Cauchy-Schwarz inequality, Fubini's theorem and the mean integrated value theorem, for $\tau_2 \geq \tau_1$,

$$\begin{aligned}\mathbb{E}\Big(V_{\tau_1} - V_{\tau_2}\Big)^2 &= \mathbb{E}\Big(\int_{\tau_1}^{\tau_2} \dot{V}_\tau\, d\tau \Big)^2 \leq \mathbb{E}(\tau_2 - \tau_1) \int_{\tau_1}^{\tau_2} \dot{V}_\tau^2\, d\tau \\ &= (\tau_2 - \tau_1) \int_{\tau_1}^{\tau_2} \mathbb{E}\dot{V}_\tau\, d\tau \leq (\tau_2 - \tau_1)^2 \sup_{\tau \in [\tau_1, \tau_2]} \mathbb{E}\dot{V}_\tau^2\, d\tau.\end{aligned}$$

□

3.6.7 Lemmas supporting the coverage results

Lemma 3.34. For $\tau \geq 1/n$ and $Y^n \sim N_n(0, I_n)$, set $H_n(\tau) = \mathbb{E}(\|\theta - \hat{\theta}(\tau)\|_2^2 \,|\, \tau, Y^n) = \sum_{i=1}^n \mathrm{var}(\theta \,|\, \tau, Y_i)$. Then for any $C > 0$, as $\tau \to 0$,

$$\sup_{t \in [C^{-1}\tau, C\tau]} \frac{1}{n\tau\check{\zeta}_\tau} \Big| H_n(t) - \mathbb{E}_0 H_n(t) \Big| \xrightarrow{P} 0.$$

Proof. Set $T = [C^{-1}\tau, C\tau]$. In view of Corollary 2.2.5 of Van der Vaart and Wellner (1996) (applied with $\psi(x) = x^2$) it is sufficient to show that $\operatorname{var}_0\left(H_n(t)/n\tau\zeta_\tau\right) \to 0$ for some $t \in T$, and

$$\int_0^{\operatorname{diam}_n} \sqrt{N(\varepsilon, T, d_n)}\, d\varepsilon = o(1), \tag{3.44}$$

where d_n is the intrinsic metric defined by its square $d_n^2(\tau_1,\tau_2) = (n\tau\zeta_\tau)^{-2}\operatorname{var}_0\Big(H_n(\tau_1) - H_n(\tau_2)\Big)$, diam_n is the diameter of the interval T with respect to the metric d_n, and $N(\varepsilon, A, d_n)$ is the covering number of the set A with ε radius balls with respect to the metric d_n.

In view of Lemma 3.8,

$$\operatorname{var}_0\left(H_n(\tau)/(n\tau\zeta_\tau)\right) \lesssim (n\tau\zeta_\tau)^{-1} \to 0.$$

Combining this with the triangle inequality and the fact that $\tau\zeta_\tau \asymp t\zeta_t$ for every $t \in T$, we also see that the diameter diam_n is bounded from above by a multiple of $1/(n\tau\zeta_\tau)^{1/2}$.

Since $d_n(\tau_1,\tau_2) \lesssim |\tau_2 - \tau_1|\tau^{-3/2}n^{-1/2}$, by Lemma 3.35, the covering number of the interval T with balls of radius ε is bounded by a multiple of $\varepsilon^{-1}/(n\tau)^{1/2}$. Hence the integral of the entropy is bounded by

$$\int_0^{\operatorname{diam}_n} \sqrt{N(\varepsilon, T, d_n)}\, d\varepsilon \lesssim (n\tau)^{-1/4} \int_0^{1/(n\tau\zeta_\tau)^{1/2}} \varepsilon^{-1/2}\, d\varepsilon \lesssim (n\tau)^{-1/2}\zeta_\tau^{-1/4} \to 0.$$

□

Lemma 3.35. For $Y_i \sim N(0,1)$, and $1/n \le \tau_1 < \tau_2 \le 1/2$,

$$\mathbb{E}_0\left(\operatorname{var}(\theta_i \mid Y_i, \tau_1) - \operatorname{var}(\theta_i \mid Y_i, \tau_2)\right)^2 \lesssim (\tau_2 - \tau_1)^2\tau_1^{-1}\zeta_{\tau_1}^2.$$

Proof. Differentiating the left side of (3.26) with respect to τ and applying Lemma 3.33 we see that the left side of the lemma is bounded above by $|\tau_1 - \tau_2|^2$ times

$$\begin{aligned}
&\sup_{\tau\in[\tau_1,\tau_2]} \mathbb{E}_0\Big[Y_i^2\frac{\dot I_{3/2}}{I_{-1/2}} - 2Y_i^2\frac{\dot I_{1/2}I_{1/2}}{I_{-1/2}^2} + \frac{\dot I_{1/2}}{I_{-1/2}} - \frac{\dot I_{-1/2}}{I_{-1/2}}\Big[Y_i^2\frac{I_{3/2}}{I_{-1/2}} - 2Y_i^2\frac{I_{1/2}^2}{I_{-1/2}^2} + \frac{I_{1/2}}{I_{-1/2}}\Big]\Big]^2 \\
&= \sup_{\tau\in[\tau_1,\tau_2]} \mathbb{E}_0\Big[Y_i^2\frac{\dot I_{3/2}}{I_{-1/2}} - 2Y_i^2\frac{\dot I_{1/2}I_{1/2}}{I_{-1/2}^2} + \frac{\dot I_{1/2}}{I_{-1/2}} + \frac{\dot I_{-1/2}}{I_{-1/2}}\Big[m_\tau(Y_i) + Y_i^2\Big[\frac{2I_{1/2}^2}{I_{-1/2}^2} - \frac{I_{1/2}}{I_{-1/2}}\Big]\Big]\Big]^2,
\end{aligned}$$

in view of (3.37). Here $\dot I_k$ denotes the partial derivative of I_k with respect to τ, and the argument Y_i of I_k and $\dot I_k$ has been omitted. In view of Lemma 3.25 and (3.42), the right hand side of the preceding display is further bounded above by a multiple of

$$\sup_{\tau\in[\tau_1,\tau_2]}\Big[\tau^2 + \tau^2\mathbb{E}_0Y_i^4 + \tau^{-2}\mathbb{E}_0m_\tau(Y_i)^2 + \tau^{-2}\mathbb{E}_0Y_i^4\frac{I_{1/2}^2}{I_{-1/2}^2}\Big].$$

The first two terms inside the square brackets are uniformly bounded, the third one is of order $o(\tau^{-1}\zeta_\tau^{-2})$ as $\tau \to 0$ in view of Lemma 3.26, and is uniformly bounded, by Lemma 3.28.

It remains to deal with the last term. By Lemmas 3.30 and 3.31 the quotient $I_{1/2}/I_{-1/2}$ is bounded by a constant for $|y| \geq \kappa_\tau$, and by a multiple of $\tau e^{y^2/2}/y^2$, otherwise. Therefore,

$$\int_{|y|\geq\kappa_\tau} y^4 \frac{I_{1/2}^2}{I_{-1/2}^2}\varphi(y)\,dy \lesssim \int_{\kappa_\tau}^{\infty} y^4 e^{-y^2/2}\,dy \lesssim e^{-\kappa_\tau^2/2}\kappa_\tau^3 \lesssim \tau\zeta_\tau,$$
$$\int_{|y|\leq\kappa_\tau} y^4 \frac{I_{1/2}^2}{I_{-1/2}^2}\varphi(y)\,dy \lesssim \tau^2 \int_0^{\kappa_\tau} e^{y^2/2}\,dy \lesssim \tau^2\kappa_\tau^{-1}e^{\kappa_\tau^2/2} \lesssim \tau\zeta_\tau.$$

This concludes the proof. □

4

Bayesian community detection

Abstract
We introduce a Bayesian estimator of the underlying class structure in the stochastic block model, when the number of classes is known. The estimator is the posterior mode corresponding to a Dirichlet prior on the class proportions, a generalized Bernoulli prior on the class labels, and a beta prior on the edge probabilities. We show that this estimator is strongly consistent when the expected degree is at least of order $\log^2 n$*, where* n *is the number of nodes in the network.*

4.1 Introduction

The stochastic block model (SBM) (Holland et al., 1983) is a model for network data in which individual nodes are considered members of classes or communities, and the probability of a connection occurring between two individuals depends solely on their class membership. It has been applied to social, biological and communication networks, for example in Park and Bader (2012), Bickel and Chen (2009) and Snijders and Nowicki (1997) amongst many others. There are many extensions of the SBM for various applications, including the degree-corrected SBM (Karrer and Newman, 2011; Zhao et al., 2012) which accounts for possible heterogeneity among nodes within the same class, and the mixed-membership SBM (Airoldi et al., 2008), in which the assumption that the classes are disjoint is removed. These extensions allow for additional modelling flexibility.

Two main SBM research directions are the recovery of the class labels (*community*

This chapter has been submitted as: S.L. van der Pas and A.W. van der Vaart. Bayesian community detection. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

detection) and recovery of the remaining model parameters, consisting of the probability vector generating the class labels, and the class-dependent probabilities of creating an edge between nodes. In this paper, we focus on community detection, noting that once strong consistency of a community detection method has been established, consistency of the natural plug-in estimators for the remaining parameters follows directly by results in (Channarond et al., 2012).

A large number of methods for recovering the class labels has been proposed. Those most closely related to this work are the modularities. Newman and Girvan (2004) introduced the term *modularity* for 'a measure of the quality of a particular division of a network'. They described one such measure for models in which edges are more likely to occur within classes than between classes, in which case there is a community structure in the colloquial sense, although the SBM does not require this assumption. Bickel and Chen (2009) studied more general modularities, defining them as functions of the number of connections between all combinations of classes and the proportion of nodes placed in each class. They introduced the likelihood modularity, and provided general conditions under which modularities are consistent. Their method and theory was extended to the degree-corrected SBM by Zhao et al. (2012).

Spectral methods for community detection have gained in popularity, and refined results on error bounds are now available for the SBM and extensions of the SBM, as evidenced in Rohe et al. (2011), Jin (2015), Sarkar and Bickel (2015) and Lei and Rinaldo (2015) for example. Many other algorithms have been introduced, most of them currently lacking formal proofs of consistency. A notable exception is the Largest Gaps algorithm (Channarond et al., 2012), which only takes the degree of each node as its input, and is strongly consistent under a separability condition.

A Bayesian approach towards recovering the class assignments in the SBM was first suggested by Snijders and Nowicki (1997), motivated by computational advantages of Gibbs sampling over maximum likelihood estimation. They considered two classes and proposed uniform priors on the class proportions and the edge probabilities. This approach was extended in (Nowicki and Snijders, 2001) to allow for more classes, with a Dirichlet prior on the class proportions and beta priors on the edge probabilities. Hofman and Wiggins (2008) described a similar Bayesian approach for a special case of the SBM and suggested a variational approach to overcome the computational issues associated with maximizing over all possible class assignments.

Bayesian methods for the SBM have barely been studied from a theoretical point of view, although recent results for parameter recovery by Pati and Bhattacharya (2015), for detecting the number of communites by Hayashi et al. (2016) and for an empirical Bayes approach to community detection by Suwan et al. (2016) are encouraging. In this work, we provide theoretical results on community detection, establishing that the Bayesian posterior mode is strongly consistent for the class labels if the expected degree is at least of order $\log^2 n$, where n is the number of nodes. This is proven by relating the posterior mode to the maximizer of the likelihood modularity of Bickel and Chen (2009). The likelihood modularity has been claimed to be strongly consistent under the weaker assumption that the expected degree is of larger order than $\log n$ (Bickel and Chen, 2009; Bickel et al., 2015; Zhao et al., 2012). However, their proof assumes that the likelihood modularity is globally Lipschitz, while it is only locally so. The Bayesian method is based on a combination

of likelihood and prior, and for this reason the proof of our main theorem, Theorem 4.3, runs into a similar problem. We were able to resolve this only under the slightly stronger assumption that the expected degree is of larger order than $(\log n)^2$. The literature on other methods for community detection shows that the order $\log n$ is sufficient for consistent detection. However, these results are usually obtained under additional assumptions such asăa restriction to two classes or an ordering of the connection probabilities, and their implications for the likelihood or Bayesian modularities is unclear. We discuss this and the relevant literature further following the statement of our main result in Section 4.3.5.

This paper is organized as follows. We introduce the SBM and the associated notation in Section 4.2. Our main results are in Section 4.3, where we describe the prior and the link with the likelihood modularity, present the consistency results and discuss the underlying assumptions, especially those on the expected degree. The method is illustrated on a data set in Section 4.4, and we conclude with a Discussion in Section 4.5. All proofs are given in the Appendix.

4.2 The stochastic block model

We introduce the notation and generative model for the SBM with $K \in \{1, 2, \ldots\}$ classes. Consider an undirected random graph with n nodes, numbered $1, 2, \ldots, n$, and edges encoded by the $n \times n$ symmetric adjacency matrix (A_{ij}), with entries in $\{0, 1\}$. Thus $A_{ij} = A_{ji}$ is equal to 1 or 0 if the nodes i and j are or are not connected by an edge, respectively. Self-loops are not allowed, so $A_{ii} = 0$ for $i = 1, \ldots, n$. The generative model for the random graph is:

1. The nodes are randomly labeled with i.i.d. variables $Z_1, \ldots, Z_n$, taking values in a finite set $\{1, \ldots, K\}$, according to probabilities $\pi = (\pi_1, \ldots, \pi_K)$.

2. Given $Z = (Z_1, \ldots, Z_n)$, the edges are independently generated as Bernoulli variables with $\mathbb{P}(A_{ij} = 1 \mid Z) = P_{Z_i, Z_j}$, for $i < j$, for a given $K \times K$ symmetric matrix $P = (P_{ab})$.

The probability vector π is considered fixed, but unknown. Although this is not visible in the notation, the matrix P may change with n, a case of particular interest being that P tends to zero, which gives a sparse graph. The order of magnitude of $\|P\|_\infty = \max_{a,b} P_{ab}$ is the same as the order of magnitude of $\rho_n = \sum_{a,b} \pi_a \pi_b P_{ab}$, the probability of there being an edge between two randomly selected nodes. The *expected degree* of a randomly selected node is $\lambda_n = (n-1)\rho_n$, and twice the expected total number of edges in the network is $\mu_n = n(n-1)\rho_n$.

The likelihood for the model is given by

$$\prod_{i<j} P_{Z_i Z_j}^{A_{ij}} (1 - P_{Z_i Z_j})^{1 - A_{ij}} \prod_i \pi_{Z_i} = \prod_{a \le b} P_{ab}^{O_{ab}(Z)} (1 - P_{ab})^{n_{ab}(Z) - O_{ab}(Z)} \prod_a \pi_a^{n_a(Z)}, \quad (4.1)$$

where $O_{ab}(Z)$ is the number of edges between nodes labelled a and b by the labelling Z, $n_{ab}(Z)$ is the maximum number of edges that can be created between nodes labelled a and b, and $n_a(Z)$ is the number of nodes labelled a, and a and b range over $\{1, 2, \ldots, K\}$.

More formally, for a given labelling $e = (e_1, \dots, e_n) \in \{1, \dots, K\}^n$ of nodes, and class labels $a, b \in \{1, \dots, K\}$, we define

$$O_{ab}(e) = \begin{cases} \sum_{i,j} A_{ij} \mathbf{1}_{\{e_i=a, e_j=b\}}, & a \neq b, \\ \sum_{i<j} A_{ij} \mathbf{1}_{\{e_i=a, e_j=b\}}, & a = b, \end{cases}$$

$$n_{ab}(e) = \begin{cases} n_a(e) n_b(e), & a \neq b, \\ \frac{1}{2} n_a(e)(n_a(e) - 1), & a = b, \end{cases}$$

$$n_a(e) = \sum_{i=1}^{n} \mathbf{1}_{\{e_i=a\}}.$$

Since the matrix A is symmetric with zero diagonal by assumption, for $a \neq b$ the variable $O_{ab}(e)$ can also be written as $\sum_{i<j} A_{ij}[\mathbf{1}_{\{e_i=a, e_j=b\}} + \mathbf{1}_{\{e_j=a, e_i=b\}}]$, which explains the different appearances of the diagonal and off-diagonal entries. The numbers $n_{ab}(e)$ are equal to the numbers $O_{ab}(e)$ when all A_{ij} are equal to 1. We collect the variables $O_{ab}(e)$ and $n_{ab}(e)$ in $K \times K$ matrices $O(e)$ and $n(e)$.

Now consider the $K \times K$ probability matrix $R(e, c)$ and K probability vector $f(e)$ with entries

$$R_{ab}(e, c) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{e_i=a, c_i=b\}}, \qquad f_a(e) = \frac{n_a(e)}{n}. \tag{4.2}$$

The row sums of $R(e, c)$ are equal to $R(e, c)\mathbf{1} = f(e)$, while the column sums are equal to $\mathbf{1}^T R(e, c) = f(c)^T$. Thus, the matrix $R(e, c)$ can be seen as a coupling of the marginal probability vectors $f(e)$ and $f(c)$. If $e = c$, then it is diagonal with diagonal $f(c) = f(e)$. More generally, the matrix can be viewed as measuring the discrepancy between labellings e and c. This can be precisely measured as half the L_1-distance of $R(e, c)$ to its diagonal, as evidenced by Lemma 4.1, which is noted in Bickel and Chen (2009).

For a vector v we denote by $\mathrm{Diag}(v)$ the diagonal matrix with diagonal v, and for a matrix M we denote its diagonal by $\mathrm{diag}\,(M)$.

Lemma 4.1. *For every labelling c, e in the K-class stochastic block model:*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{c_i \neq e_i\}} = \tfrac{1}{2} \|\mathrm{Diag}(f(c)) - R(e, c)\|_1.$$

Proof. The diagonal of $R(e, c)$ gives the fractions of labels on which c and e agree. Hence the left side of the lemma is $1 - \sum_a R_{aa}(e, c) = \sum_a (f_a(c) - R_{aa}(c))$. The elements of both $K \times K$ matrices $\mathrm{Diag}(f(c))$ and $R(e, c)$ can be viewed as probabilities that add up to 1. Thus the sum of the differences of the diagonal elements is minus the sum of the differences of the off-diagonal elements. Because $f_a(c) \geq R_{aa}(e, c)$ for every a, we have $\sum_a (f_a(c) - R_{aa}(e, c)) = \sum_a |f_a(c) - R_{aa}(e, c)|$. Similarly the off-diagonal elements of $\mathrm{Diag}(f(c))$, which are zero, are smaller than the off-diagonal elements of $R(e, c)$ and hence we can add absolute values. Thus the sum over the diagonal is half the sum of the absolute values of all terms in $\mathrm{Diag}(f(c)) - R(e, c)$. □

4.3 Bayesian approach to community detection

Our main results are presented in this section. We first discuss the choice of prior in Section 4.3.1, and define the estimator, in Section 4.3.2. The resulting Bayesian modularity is closely related to the likelihood modularity of Bickel and Chen (2009). The relationship is clarified in Section 4.3.3. We briefly consider the issue of identifiability in the SBM in Section 4.3.4, and conclude with our main theorem on the strong consistency of the Bayesian modularity in Section 4.3.5.

4.3.1 The prior

We adopt the Bayesian approach of Nowicki and Snijders (2001). We put prior distributions on the parameters of the stochastic block model with K known, the vector π and the matrix P, yielding a joint probability distribution of (A, Z, π, P). Next we marginalize over π and P as in McDaid et al. (2013), leading to a joint distribution of (A, Z). Finally we "estimate" the unobserved vector Z by the posterior mode of the conditional distribution of Z given A. From a frequentist point of view this means that Z is treated as a parameter of the problem, equipped with a hierarchical prior that chooses first π and then Z. Accordingly we shall change notation from Z to e, reserving Z for the frequentist description of the stochastic block model in Section 4.2.

The prior on π is a Dirichlet, and independently the P_{ab} for $a \leq b$ receive independent beta priors:

$$\begin{aligned} \pi &\sim \mathrm{Dir}(\alpha, \ldots, \alpha), \\ P_{ab} &\overset{i.i.d.}{\sim} \mathrm{Beta}(\beta_1, \beta_2), \quad 1 \leq a \leq b \leq K. \end{aligned}$$

This is essentially the same set-up as in Nowicki and Snijders (2001) and McDaid et al. (2013), except that we use a more flexible $\mathrm{Beta}(\beta_1, \beta_2)$ instead of a uniform prior on the P_{ab}. We assume $\alpha, \beta_1, \beta_2 > 0$.

We complete the Bayesian model by specifying class labels $e = (e_1, \ldots, e_n)$ and edges $A = (A_{ij} : i < j)$ through

$$\begin{aligned} e_i \mid \pi, P &\overset{i.i.d.}{\sim} \pi, \quad 1 \leq i \leq n, \\ A_{ij} \mid \pi, P, e &\overset{ind.}{\sim} \mathrm{Bernoulli}(P_{e_i, e_j}), \quad 1 \leq i < j \leq n. \end{aligned}$$

Abusing notation we write $p(e)$, $p(A \mid e)$ and $p(e \mid A)$ for marginal and conditional probability density functions.

4.3.2 The Bayesian modularity

The Bayesian estimator of the class labels will be the posterior mode, that is:

$$\widehat{e} = \operatorname*{argmax}_e p(e \mid A).$$

The posterior mode can be interpreted as a modularity-based estimator in the sense of Bickel and Chen (2009), in that it maximizes a function that only depends on the $O_{ab}(e)$ and

the $n_a(e)$. This can be seen from the joint density of (A, e), which is found by marginalizing the likelihood (4.1) over π and P. The conjugacy between the multinomial and Dirichlet distributions gives the marginal density of the class assignment e as:

$$p(e) = \int_{S_K} \prod_a \pi_a^{n_a(e)} \frac{\prod_a \pi_a^{\alpha-1}}{D(\alpha)}\, d\pi = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K \Gamma(n+\alpha K)} \prod_a \Gamma(n_a(e)+\alpha). \tag{4.3}$$

Here the integral is relative to the Lebesgue measure on the K-dimensional unit simplex and $D(\alpha) = \Gamma(\alpha)^K/\Gamma(K\alpha)$ is the norming constant for the Dirichlet density. Similarly the conjugacy between the Bernoulli and Beta distributions gives the marginal conditional density of A given e as:

$$\begin{aligned} p(A \mid e) &= \int_{[0,1]^{K(K+1)/2}} \prod_{a\le b} P_{ab}^{O_{ab}(e)} (1-P_{ab})^{n_{ab}(e)-O_{ab}(e)} \prod_{a\le b} \frac{P_{ab}^{\beta_1-1}(1-P_{ab})^{\beta_2-1}}{B(\beta_1,\beta_2)}\, dP \\ &= \prod_{a\le b} \frac{1}{B(\beta_1,\beta_2)} B(O_{ab}(e)+\beta_1, n_{ab}(e)-O_{ab}(e)+\beta_2), \end{aligned} \tag{4.4}$$

where $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the beta-function. The joint density of A and e is given by the product of (4.3) and (4.4), and n^{-2} times its logarithm is up to a constant that is free of e equal to

$$Q_B(e) = \frac{1}{n^2} \sum_{1\le a\le b\le K} \log B(O_{ab}(e)+\beta_1, n_{ab}(e)-O_{ab}(e)+\beta_2) + \frac{1}{n^2}\sum_{a=1}^{K} \log\Gamma(n_a(e)+\alpha).$$

This is a modularity in the sense of Bickel and Chen (2009), which we define as the *Bayesian modularity*. As $p(e \mid A)$ is proportional to $p(e, A)$, the posterior mode is equal to the class assignment that maximizes the Bayesian modularity, so the Bayesian estimator is equal to:

$$\widehat{e} = \operatorname*{argmax}_e Q_B(e). \tag{4.5}$$

4.3.3 Similarity to the likelihood modularity

The Bayesian modularity $Q_B(e)$ consists of a two parts, originating from the likelihood and the prior on the classification, respectively. The first part is close to the *likelihood modularity* given by

$$Q_{ML}(e) = \frac{1}{n^2} \sum_{1\le a\le b\le K} n_{ab}(e)\, \tau\Big(\frac{O_{ab}(e)}{n_{ab}(e)}\Big),$$

where $\tau(x) = x\log x + (1-x)\log(1-x)$. This criterion, obtained in Bickel and Chen (2009), results from replacing in the log conditional likelihood of A given e (the logarithm of (4.1) with Z replaced by e and discarding the term involving the parameters π_a) the parameters P_{ab} by their maximum likelihood estimators $\hat{P}_{ab} = O_{ab}(e)/n_{ab}(e)$. In other words, the parameters are *profiled out* rather than integrated out as for the Bayesian modularity. The corresponding estimator

$$\widehat{e}_{ML} = \operatorname*{argmax}_e Q_{ML}(e)$$

is consistent, and hence one may hope that the Bayesian estimator can be proved consistent by showing that the Bayesian and likelihood modularities are close. This will indeed be our line of approach, but the execution must be done with care. For instance, the second, prior part of the Bayesian modularity does play a role in the proof of strong consistency, although it is negligible when proving weak consistency.

The following lemma links the Bayesian and likelihood modularities.

Lemma 4.2. *There exists a constant C such that, for $\mathcal{E} = \{1,\dots,K\}^n$ the set of all possible labellings:*

$$\max_{e\in\mathcal{E}}\Big|Q_B(e) - Q_{ML}(e) - Q_P(e)\Big| \le \frac{C\log n}{n^2},$$

for

$$Q_P(e) = \frac{1}{n^2}\sum_{a:n_a(e)+\lfloor\alpha\rfloor\ge 2} n_a(e)\log(n_a(e)) - \frac{1}{n}.$$

Consequently $\max_{e\in\mathcal{E}}\big|Q_B(e) - Q_{ML}(e)\big| = O\big(\log n/n\big)$.

4.3.4 Identifiability and consistency

A classification $\widehat{e}$ is said to be *weakly consistent* if the fraction of misclassified nodes tends to zero (partial recovery), and *strongly consistent* if the probability of misclassifying any of the nodes tends to zero (exact recovery). In defining consistency in a precise manner, the complication of the possible unidentifiability of the labels needs to be dealt with. From the observed data A we can at best recover the partition of the n nodes in the K classes with equal labels Z_i, but not the values $Z_1,\dots,Z_n$ of the labels, in the set $\{1,2,\dots,K\}$, attached to the classes. Thus consistency will be up to a permutation of labels.

To make this precise define, for a given permutation $(1,\dots,K) \to (\sigma(1),\dots,\sigma(K))$, the *permutation matrix* P_σ as the matrix with rows

$$\begin{matrix} e_{\sigma(1)}^T \\ \vdots \\ e_{\sigma(K)}^T, \end{matrix}$$

for $e_1,\dots,e_K$ the unit vectors in $\mathbb{R}^K$. Then pre-multiplication of a matrix by P_σ permutes the rows, and post-multiplication by P_σ^T the columns: $P_\sigma R$ is the matrix with jth row equal to the $\sigma(j)$th row of R, and RP_σ^T is the matrix with jth column the $\sigma(j)$th column of R. Thus $P_\sigma R(e,Z)$ is the matrix that would result if we would permute the labels of the classes of the assignment e, and $P_\sigma P P_\sigma^T$ and $P_\sigma R(e,Z)P_\sigma^T$ are the matrices that would result if we would relabel the classes throughout. Since we cannot recover the labels, the matrix $P_\sigma R(e,Z)$ is just as good or bad as $R(e,Z)$ for measuring discrepancy between a labelling e and the true labelling Z; furthermore, nothing should change if we choose different names for the classes.

Thus, taking into account the unidentifiability of the labels, by Lemma 4.1, an estimator $\widehat{e}$ is *weakly consistent* if

$$\|P_\sigma R(\widehat{e}, Z) - \mathrm{Diag}(f(Z))\|_1 \to 0,$$

for some permutation matrix P_σ. The classification $\widehat{e}$ is said to be *strongly consistent* if

$$\mathbb{P}(P_\sigma R(\widehat{e}, Z) = \mathrm{Diag}(f(Z))) \to 1,$$

for some permutation matrix P_σ.

The permutation matrix P_σ is for large n uniquely defined: if $\|(P_\sigma)_j R - \mathrm{Diag}(\pi)\|_1 \leq \min_a \pi_a$, for $j = 1, 2$, then $(P_\sigma)_1 = (P_\sigma)_2$. This follows because the assumption implies that $\|(P_\sigma)_1^{-1}\mathrm{Diag}(\pi) - (P_\sigma)_2^{-1}\mathrm{Diag}(\pi)\|_1 \leq 2\min_a \pi_a$, by the triangle inequality and the fact that the L_1-norm is invariant under permutations. Furthermore, for $P_\sigma = (P_\sigma)_2(P_\sigma)_1^{-1}$ the left side is $\|P_\sigma \mathrm{Diag}(\pi) - \mathrm{Diag}(\pi)\|_1$, which is at least two times the sum of the two smallest coordinates of π if $P_\sigma \neq I$.

A necessary requirement for consistency is that the classes can be recovered from the likelihood, i.e. the model parameters must be identifiable. If π has strictly positive coordinates, so that all labels will appear in the data eventually, then as explained in Bickel and Chen (2009) an appropriate condition is that P does not have two identical rows. If $\pi_a = 0$ for some a, then class a will never be consumed; the identifiability condition should then be imposed after deleting the ath column from P. Thus, we call the pair (P, π) *identifiable* if the rows of P are different after removing the columns corresponding to zero coordinates of π. Throughout we assume that P is symmetric.

4.3.5 Consistency results and assumptions

We are now ready to present our results on consistency for the Bayesian maximum a posteriori (MAP) estimator (4.5). Theorem 4.3 shows strong consistency of the Bayesian estimator if $\lambda_n \gg (\log n)^2$. The proof rests on a proof of weak consistency under similar conditions, stated in the appendix as Theorem 4.4.

Recall that $\rho_n = \sum_{a,b} \pi_a \pi_b P_{ab}$ is the probability of a new edge, and $\lambda_n = (n-1)\rho_n$ is the expected degree of a node.

Theorem 4.3 (strong consistency). *(i) If (P, π) is fixed and identifiable with $0 < P < 1$ and $\pi > 0$ then the MAP classifier $\widehat{e} = \arg\max_e Q_B(e)$ is strongly consistent.*

(ii) If $P = \rho_n S$, where (S, π) is fixed and identifiable with $S > 0$ and $\pi > 0$, then the MAP classifier $\widehat{e} = \arg\max_e Q_B(e)$ is strongly consistent if $\lambda_n \gg (\log n)^2$.

The theorem distinguishes two cases: i is the *dense* case, while ii is the *sparse* case. The second is the most interesting of the two, as it touches on the question how much information is required to recover the underlying community structure. Much recent research effort has gone into determining detection and computational boundaries, in particular for special cases of the SBM with $K = 2$ (see e.g. Mossel et al. (2012), Chen and Xu (2014), Abbe et al. (2014) and Zhang and Zhou (2015)).

Weakly consistent estimation of the class labels for an arbitrary, but known, number of classes is possible under the assumption $\lambda_n \gg \log n$, as this was shown to hold for

spectral clustering by Lei and Rinaldo (2015). *Strong* consistency of maximum likelihood was shown to hold in the special cases of planted bisection ($K = 2$ and equal community sizes) and planted clustering (equal community sizes and P_{ab} can take two values) by Abbe et al. (2014); Chen and Xu (2014), again under the assumption $\lambda_n \gg \log n$. Gao et al. (2015) and Gao et al. (2016) achieve optimality in different senses, under assumptions on the average within-community and between-community edge probabilities; Gao et al. (2015) introduce a two-stage procedure which achieves the optimal proportion of misclassified nodes in a special case where P_{ab} can only take two values, while Gao et al. (2016) obtain minimax rates for the proportion of misclassified nodes in the degree corrected SBM.

Strong consistency of the likelihood modularity for an arbitrary number of classes K has been claimed under the same assumption $\lambda_n \gg \log n$ (Bickel and Chen, 2009), and those results have been extended to the degree-corrected SBM (Zhao et al., 2012). However, these results were obtained by application of an abstract theorem to the special case of the likelihood modularity, which would require the function $\tau(x) = x \log x + (1 - x) \log(1 - x)$, or the function $\sigma(x) = x \log x$, to be globally Lipschitz. As τ and σ are only locally Lipschitz, it is still unclear whether $\lambda_n \gg \log n$ is a sufficient condition for either weakly or strongly consistent estimation by maximum likelihood. From our proof of Theorem 4.3, which proceeds by comparing the Bayesian modularity to the likelihood modularity, it immediately follows that $\lambda_n \gg (\log n)^2$ is certainly sufficient. Given weak consistency the problem can be reduced to a neighbourhood of the true parameter on which the Lipschitz condition is reasonable. However, it is precisely our proof of weak consistency that needs the additional $\log n$ factor.

The Largest Gaps algorithm of Channarond et al. (2012) is strongly consistent provided that $\min_{a \neq b} |\sum_{k=1}^{K} \alpha_k (P_{ak} - P_{bk})|$ is at least of order $\sqrt{\log n / n}$, implying that at least one of the P_{ab} is of the same order, and thus $\lambda_n \gg \sqrt{n \log n}$. This much stronger condition is not surprising, as the Largest Gaps algorithm only uses the degree of a node and does not take into account any finer information on the group structure, such as the information contained in the O_{ab}.

To the best of our knowledge, for $K > 2$, it remains to be shown that $\lambda \gg \log n$ is sufficient for strong consistency of any community detection method for the general SBM. For the minimax rate for the proportion of misclustered nodes in community detection, when only classes of sizes proportional to n are considered, a phase transition when going from the case $K = 2$ to $K \geq 3$ was observed by Zhang and Zhou (2015). Their results show that if $K = 2$, communities of the same size are most difficult to distinguish, while if $K \geq 3$, small communities are harder to discover. This shift in the nature of the communities that are harder to detect may be what has been preventing a general strong consistency result under the assumption $\lambda_n \gg \log n$ so far.

4.4 Application to the karate club data set

Some options for implementing the Bayesian modularity are given in Section 4.4.1, after which the results of applying the Bayesian and likelihood modularities to the well-studied karate club data of Zachary (1977) are discussed in Section 4.4.2.

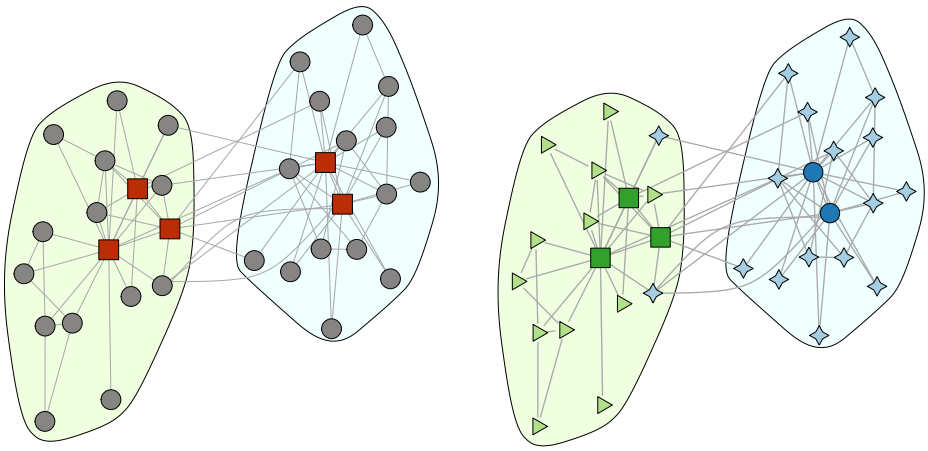

Figure 4.1: Communities detected by the Bayesian modularity when $K = 2$ (left) and $K = 4$ (right), with $\alpha = \beta_1 = \beta_2 = 1/2$. The polygons contain the two groups the karate club was split into; the left one is Mr. Hi's club, the right one is the Officers' club. The shapes of the nodes represent the communities selected by the modularities. Figure made using the igraph package (Csardi and Nepusz, 2006).

4.4.1 Implementation

Two recent works explicitly discuss implementation of Bayesian methods for the SBM. McDaid et al. (2013) followed the approach of Nowicki and Snijders (2001) and added a Poisson prior on K. After marginalizing over π and P, they employ an allocation sampler to sample from the joint density of K and z given A, and use the posterior mode to estimate K. Their algorithm can scale to networks with approximately ten thousand nodes and ten million edges. Côme and Latouche (2014), claiming that the algorithm of McDaid et al. (2013) suffers from poor mixing properties, propose a greedy inference algorithm for the same problem. For the karate club data in Section 4.4.2, the network was small enough that a tabu search (Glover, 1989), run for a number of different initial configurations, yielded good results. We used $\alpha = 1/2$ for the Dirichlet prior, and $\beta_1 = \beta_2 = 1/2$ for the beta prior.

4.4.2 Karate club

Zachary (1977) described a karate club which split into two clubs after a conflict over the price of the karate lessons. The new club was led by Mr. Hi, the karate teacher of the original club, while the remainder of the old club stayed under the former Officers' rule. The data consists of an adjacency matrix for those 34 individuals who interacted with other club members outside club meetings and classes. Each of these individuals' affiliations after the conflict is known.

The communities selected by the Bayesian modularity for $K = 2$ and $K = 4$ are given in Figure 4.1. In both instances, the tabu search led to nearly the same solution for both the

Bayesian and likelihood modularities, only differing at one node for $K = 4$, which is not surprising in light of Lemma 4.2. For $K = 2$, the results of Bickel and Chen (2009) for this data set are recovered. For $K = 4$, the partition in Figure 4.1 yields a higher value of the likelihood modularity than the partition into four classes found by Bickel and Chen (2009), and an even higher value is obtained by switching club member 20 to the second-largest class. This discrepancy is likely due to the heuristic nature of the tabu search algorithm, and for the same reason, it may be the case that improvement over the partitions found by the Bayesian modularity in Figure 4.1 are possible.

For $K = 2$, the communities found by the algorithms do not correspond in the slightest to the two karate clubs, instead grouping the nodes with the highest degrees, corresponding to Mr. Hi, the president of the original club, and their closest supporters, together. Incidentally, this partition is the same as the one returned by the Largest Gaps algorithm of Channarond et al. (2012), which solely uses the degrees of the nodes and discards all other information.

These bad results are no reason to shelve the Bayesian and likelihood modularities, as there is no reason to believe that the two karate clubs form communities in the sense of the stochastic block model. Mr. Hi and the club's president are clear outliers within their groups, and neither of the algorithms were designed to be robust to such a phenomenon. The communities selected by the modularities are communities in the sense that they form connections within and between the groups in a similar fashion. This sense does not correspond to the social notion of a community in this setting.

The results for four classes unify the social and stochastic senses of community. The prominent members of each of the new clubs are placed into two separate, small, communities. The other members are classified nearly perfectly, with two exceptions. However, one of those exceptional individuals is the only person described by Zachary (1977) as being a supporter of the club's president before the split, who joined Mr. Hi's club, making this person's affiliation up for debate. The second is described as only a weak supporter of Mr. Hi. The increased number of communities allows for some outliers within the social communities, and leads to a more detailed understanding of the dynamics within both of the groups. We essentially recover the two communities, each with a core that is more connective than the remainder of the nodes.

4.5 Discussion

An advantage of Bayesian modelling is that it does not solely result in an estimator, but in a full posterior distribution. The posterior mode studied in this paper is but one aspect of the posterior, and its good behaviour in terms of consistency is encouraging. Further study into other aspects in the posterior may prove to be fruitful. One possible research direction would be to use the posterior to *quantify uncertainty* in the estimate of the class labels. A second issue that may be resolved by the Bayesian approach is the question of estimating the number of classes, K. This remains an important open question, as noted by Bickel and Chen (2009), despite recent attempts (e.g. Saldana et al. (2014), Chen and Lei (2014) and Wang and Bickel (2015)). By introducing a prior on K, such as the Poisson-prior suggested by McDaid et al. (2013), the number of communities K can be detected by the posterior.

4.6 Proofs

After stating some repeatedly used notation, this appendix starts with the proof of Theorem 4.4, which is a theorem on weak consistency of the Bayesian modularity. It is followed by a number of supporting Lemmas, after which we proceed to the proof of Theorem 4.3, and some additional supporting Lemmas.

We write diag (P) for the diagonal of P if P is a matrix, and Diag(f) for the diagonal matrix with diagonal f if f is a vector.

4.6.1 Weak consistency

The following quantities will be used in the course of multiple proofs. The function H_P, with domain $K \times K$ probability matrices, is given by, for $\tau(u) = u \log u + (1-u)\log(1-u)$,

$$H_P(R) = \frac{1}{2}\sum_{a,b}(R\mathbf{1})_a(R\mathbf{1})_b\,\tau\left(\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\right). \tag{4.6}$$

For $\tau_0(u) = u\log(u) - u$, define

$$G_P(R) = \frac{1}{2}\sum_{a,b}(R\mathbf{1})_a(R\mathbf{1})_b\,\tau_0\Big(\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\Big).$$

The sums defining these functions are over all pairs (a,b) with $1 \le a,b \le K$, unlike the sums defining the modularities Q_B and Q_{ML}, which are restricted to $a \le b$.

Theorem 4.4 (weak consistency)**.** *(i) If (P,π) is fixed and identifiable, then the MAP classifier $\widehat{e} = \arg\max_z Q_B(e)$ is weakly consistent.*

(ii) If $P = \rho_n S$ for $\rho_n \to 0$, and (S,π) is fixed and identifiable, then the MAP classifier $\widehat{e} = \arg\max_z Q_B(e)$ is weakly consistent provided $n\rho_n \gg (\log n)^2$.

Proof. By Lemma 4.2 the Bayesian modularity Q_B is equivalent to the likelihood modularity Q_{ML} up to order $(\log n)/n$. With the notation $\widetilde{O}_{ab}(e) = O_{ab}(e)$ if $a \ne b$, and $\widetilde{O}_{ab}(e) = 2O_{ab}(e)$ if $a = b$, the likelihood modularity is in turn equivalent up to the same order to

$$\mathbb{L}(e) = \frac{1}{2n^2}\sum_{a,b} n_a(e)n_b(e)\,\tau\Big(\frac{\widetilde{O}_{ab}(e)}{n_a(e)n_b(e)}\Big). \tag{4.7}$$

Indeed the terms of $Q_{ML}(e)$ for $a < b$ are identical to the sums of the terms of $\mathbb{L}(e)$ for $a < b$ and $a > b$, while for $a = b$ the terms of $Q_{ML}(e)$ and $\mathbb{L}(e)$ differ only subtly: the first uses $n_{aa}(e) = \frac{1}{2}n_a(e)(n_a(e)-1)$, where the second uses $\frac{1}{2}n_a(e)^2$. Thus the difference is bounded in absolute value by the sum over a of (where e is suppressed from the notation)

$$\Big|\frac{n_a^2}{2n^2}\tau\Big(\frac{\widetilde{O}_{aa}}{n_a^2}\Big) - \frac{n_a\big(n_a-1\big)}{2n^2}\tau\Big(\frac{\widetilde{O}_{aa}}{n_a(n_a-1)}\Big)\Big| \le \frac{1}{2n}\|\tau\|_\infty + \frac{n_a^2}{2n^2}l\Big(\frac{\widetilde{O}_{aa}}{n_a^2(n_a-1)}\Big).$$

where $l(x) = x(1\vee\log(1/x))$, in view of Lemma 4.7. We now use that $n_a l(u/n_a) \lesssim \log n_a \le \log n$, for $0 \le u \le 1$.

Combining the preceding, we conclude that

$$\eta_{n,1} := \max_e |\mathbb{L}(e) - Q_B(e)| = O\left(\frac{\log n}{n}\right).$$

Since $Q_B(\widehat{e}) \geq Q_B(Z)$, by the definition of $\widehat{e}$, it follows that $\mathbb{L}(\widehat{e}) - \mathbb{L}(Z) \geq -2\eta_{n,1}$. The next step is to replace $\mathbb{L}$ in this equality by an asymptotic value.

For x equal to a big multiple of $(\|P\|_\infty^{1/2} \vee n^{-1/2})/n^{1/2}$, the right side of Lemma 4.5 tends to zero and hence $\max_e \left\|\widetilde{O}(e) - \mathbb{E}\big(\widetilde{O}(e) \mid Z\big)\right\|_\infty / n^2$ is of this order in probability. We also have, by Lemma 4.6:

$$\max_e \left\| \frac{1}{n^2} \mathbb{E}\big(\widetilde{O}(e) \mid Z\big) - R(e,Z)PR(e,Z)^T \right\|_\infty = \max_e \frac{1}{n} \left\| \mathrm{Diag}(R(e,Z))\, \mathrm{diag}\,(P) \right\|_\infty \to 0,$$

as each entry of $\mathrm{Diag}(R(e,Z))\,\mathrm{diag}\,(P)$ is bounded above by one. By Lemma 4.7, $\big|v\tau(x/v) - v\tau(y/v)\big| \leq l(|x-y|)$, uniformly in $v \in [0,1]$, where $l(x) = x(1 \vee \log(1/x))$. It follows that

$$\eta_{n,2} := \max_e \big|\mathbb{L}(e) - L(e)\big| = o_P\Big(l\Big(\frac{\|P\|_\infty^{1/2} \vee n^{-1/2}}{n^{1/2}}\Big)\Big),$$

for

$$L(e) = \frac{1}{2} \sum_{a,b} f_a(e) f_b(e)\, \tau\Big(\frac{(R(e,Z)PR(e,Z)^T)_{ab}}{f_a(e) f_b(e)}\Big).$$

Combining this with the preceding paragraph, we conclude that $L(\widehat{e}) \geq L(Z) - 2(\eta_{n,1} + \eta_{n,2})$.

Proof of i. For given $\delta > 0$, let $\mathcal{R}_\delta$ be the set of all probability matrices R with

$$\min_{P_\sigma} \left\| P_\sigma R - \mathrm{Diag}(R^T \mathbf{1}) \right\|_1 \geq \delta, \qquad \text{and} \qquad \min_{a:\pi_a > 0} (R^T \mathbf{1})_a \geq \delta.$$

Here the minimum is taken over the (finite) set of all permutation matrices P_σ on K labels. Furthermore, set

$$\eta := \inf_{R \in \mathcal{R}_\delta} \Big[H_P\Big(\mathrm{Diag}(R^T \mathbf{1})\Big) - H_P(R) \Big],$$

where H_P is as defined in (4.6). Because $\mathcal{R}_\delta$ is compact and the maps $R \mapsto H_P(R)$ and $R \mapsto \mathrm{Diag}(R^T \mathbf{1})$ are continuous, the infimum in the display is assumed for some $R \in \mathcal{R}_\delta$. Because no $R \in \mathcal{R}_\delta$ can be transformed into a diagonal element by permuting rows and every $R \in \mathcal{R}_\delta$ has a nonzero element in every column a with $\pi_a > 0$, Lemma 4.8 shows that $\eta_n > 0$.

Because $L(e) = H_P(R(e,Z))$ for every e, and $R(Z,Z) = \mathrm{Diag}(f(Z)) = \mathrm{Diag}(R(\widehat{e},Z)^T \mathbf{1})$, we conclude that

$$H_P(\mathrm{Diag}(R(\widehat{e},Z)^T \mathbf{1})) - H_P(R(\widehat{e},Z)) \leq 2(\eta_{n,1} + \eta_{n,2}).$$

If $2(\eta_{n,1} + \eta_{n,2})$ is smaller than η_n, then it follows that $R(\widehat{e},Z)$ cannot be contained in $\mathcal{R}_\delta$. Since $R(\widehat{e},Z)^T \mathbf{1} = f(Z) \xrightarrow{P} \pi$, by the law of large numbers, for sufficiently small $\delta > 0$ this must be because $R(\widehat{e},Z)$ fails the first requirement defining $\mathcal{R}_\delta$. That is, $\|P_\sigma R(\widehat{e},Z) -$

$\mathrm{Diag}(f(Z))\|_1 \le \delta$ for some permutation matrix P_σ. As this is true eventually for any $\delta > 0$, it follows that $\min_{P_\sigma} \|P_\sigma R(\widehat{e}, Z) - \mathrm{Diag}(\pi)\|_1 \xrightarrow{P} 0$.

Proof of ii. In view of Lemma 4.9, the number $\eta = \eta_n$, which now depends on n, is now bounded below by ρ_n times a positive number that depends on (S, π). The preceding argument goes through provided $\eta_{n,1} + \eta_{n,2}$ is of smaller order than η_n. This leads to $l\Big(\sqrt{\rho_n/n}\Big) + \log(n)/n \ll \rho_n$, or $(\rho_n/n)\log^2\Big(n/(\rho_n\|S\|_\infty)\Big) \ll \rho_n^2$. □

Lemma 4.5. *Let $\widetilde{O}_{ab}(e) = O_{ab}(e)$ if $a \neq b$, and $\widetilde{O}_{ab}(e) = 2O_{ab}(e)$ if $a = b$. For any $x > 0$,*

$$\mathbb{P}\Big(\max_e \Big\|\widetilde{O}(e) - \mathbb{E}\Big(\widetilde{O}(e) \mid Z\Big)\Big\|_\infty > xn^2\Big) \le 2K^{n+2} e^{-x^2 n^2/(8\|P\|_\infty + 4x/3)}.$$

Proof. This Lemma is adapted from Lemma 1.1 in Bickel and Chen (2009). There are K^n possible values of e and $\|\cdot\|_\infty$ is the maximum of the K^2 entries in the matrix. We use the union bound to pull these maxima out of the probability, giving the factor K^{n+2} on the right. Next it suffices to bound the tail probability of each variable

$$\widetilde{O}_{ab}(e) - \mathbb{E}\Big(\widetilde{O}_{ab}(e) \mid Z\Big) = \sum_{i,j}\Big(A_{ij} - \mathbb{E}(A_{ij} \mid Z)\Big)(\mathbf{1}\{e_i = a, e_j = b\} + \mathbf{1}\{e_i = b, e_j = a\}).$$

The $n_{ab}(e)$ variables in this sum are conditionally independent given Z, take values in $[-2, 2]$, and have conditional mean zero given Z and conditional variance bounded by $4\,\mathrm{var}(A_{ij} \mid Z) \le 4P_{Z_i Z_j}(1 - P_{Z_i Z_j}) \le 4\|P\|_\infty$. Thus we can apply Bernstein's inequality to find that

$$\mathbb{P}\Big(\Big|\widetilde{O}_{ab}(e) - \mathbb{E}\Big(\widetilde{O}_{ab}(e) \mid Z\Big)\Big| > xn^2\Big) \le 2e^{-x^2 n^4/(8n_{ab}(e)\|P\|_\infty + 4xn^2/3)}.$$

Finally we use the crude bound $n_{ab}(e) \le n^2$ and cancel one factor n^2. □

Lemma 4.6. *Define $\widetilde{O}_{ab}(e) = O_{ab}(e)$ if $a \neq b$, and $\widetilde{O}_{ab}(e) = 2O_{ab}(e)$ if $a = b$. Then, for $R(e, Z)$ as defined in* (4.2)*,*

$$\mathbb{E}(\widetilde{O}_{ab} \mid Z) = n^2 R(e, Z) P R(e, Z)^T - n\mathrm{Diag}(R(e, Z)\,\mathrm{diag}\,(P)).$$

Proof. A similar expression, not taking into account the absence of self-loops, appears in Bickel and Chen (2009).

$$\begin{aligned}
\mathbb{E}(\widetilde{O}_{ab}(e) \mid Z = c) &= \sum_{i \neq j} P_{c_i c_j} \mathbf{1}\{e_i = a, e_j = b\} \\
&= \sum_{a', b'} P_{a'b'} \sum_{i \neq j} \mathbf{1}\{c_i = a', c_j = b'\}\mathbf{1}\{e_i = a, e_j = b\} \\
&= \sum_{a', b'} P_{a'b'} \sum_{i,j} \mathbf{1}\{c_i = a', c_j = b'\}\mathbf{1}\{e_i = a, e_j = b\} - \delta_{ab} \sum_{a'} P_{a'a'} \mathbf{1}\{c_i = a'\}\mathbf{1}\{e_i = a\} \\
&= n^2 \sum_{a', b'} P_{a'b'} R_{aa'}(e, c) R_{bb'}(e, c) - \delta_{ab} n \sum_{a'} P_{a'a'} R_{aa'}(e, c).
\end{aligned}$$

□

Lemma 4.7. *The function $\tau : [0,1] \to \mathbb{R}$ satisfies $|\tau(x) - \tau(y)| \le l(|x-y|)$, for $l(x) = 2x(1 \vee \log(1/x))$.*

Proof. Write the difference between $x \log x$ and $y \log y$ as $|\int_x^y (1 + \log s)\, ds|$. The function $s \mapsto 1 + \log s$ is strictly increasing on $[0,1]$ from $-\infty$ to 1 and changes sign at $s = e^{-1}$. Therefore the absolute integral is bounded above by the maximum of

$$-\int_0^{|x-y| \wedge e^{-1}} (1 + \log s)\, ds = -(|x-y| \wedge e^{-1}) \log |x-y| \wedge e^{-1}$$

and

$$\int_{1-|x-y| \vee e^{-1}}^1 (1 + \log s)\, ds \le |x-y|.$$

□

Proof of Lemma 4.2

Proof. The second assertion of the lemma follows from the first and the fact that $\max_e Q_P(e) \lesssim (\log n)/n$. It suffices to prove the first assertion.

Recall that the Bayesian modularity is given by

$$n^2 Q_B(e) = \sum_{a \le b} \log B\left(O_{ab}(e) + \tfrac{1}{2}, n_{ab}(e) - O_{ab}(e) + \tfrac{1}{2}\right) + \sum_a \log \Gamma(n_a(e) + \alpha). \tag{4.8}$$

We shall show that the first sum on the right is equivalent to $Q_{ML}(e)$, and the second sum is equivalent to $Q_P(e)$. We show this by comparing the sums defining the various modularities term by term. For clarity we shall suppress the argument e. We will repeatedly use the following bound from (Robbins, 1955): for $n \in \mathbb{N}_{\ge 1}$,

$$\Gamma(n+1) = \sqrt{2\pi} n^{n+1/2} e^{-n} e^{a_n}, \tag{4.9}$$

with $(12n+1)^{-1} \le a_n \le (12n)^{-1}$, as well as the fact that $\Gamma(s)$ is monotone increasing for $s \ge 3/2$. In addition, we will bound remainder terms by using the inequality $x \log((x+c)/x) \le c$ for $c \ge 0$ and the fact that $x \log((x-1)/x)$ is bounded for $x > 1$.

First sum of (4.8)**.**

Upper bound, case 1: $O_{ab} \ne 0$ and $n_{ab} \ne O_{ab}$

We apply (4.9):

$$\begin{aligned}
\log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) &\le \log \frac{\Gamma(O_{ab} + \lfloor \beta_1 \rfloor + 1)\Gamma(n_{ab} - O_{ab} + \lfloor \beta_2 \rfloor + 1)}{\Gamma(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor)} \\
&= O_{ab} \log\left(\frac{O_{ab} + \lfloor \beta_1 \rfloor}{n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor - 1}\right) + (n_{ab} - O_{ab}) \log\left(\frac{n_{ab} - O_{ab} + \lfloor \beta_2 \rfloor}{n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor - 1}\right) \\
&\quad + (\lfloor \beta_1 \rfloor + 1/2) \log(O_{ab} + \lfloor \beta_1 \rfloor) + (\lfloor \beta_2 \rfloor + 1/2) \log(n_{ab} - O_{ab} + \lfloor \beta_2 \rfloor) \\
&\quad - (\lfloor \beta_1 + \beta_2 \rfloor - 1/2) \log(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor - 1) + \log \sqrt{2\pi} - \lfloor \beta_1 \rfloor - \lfloor \beta_2 \rfloor + \lfloor \beta_1 + \beta_2 \rfloor - 1 \\
&\quad + \alpha_{ab} + \beta_{ab} - \gamma_{ab},
\end{aligned}$$

where α_{ab}, β_{ab} and γ_{ab} are bounded by constants. By the inequality $x \log((x+c)/x) \leq c$ for $c \geq 0$, and the fact that $x \log((x-1)/x)$ is bounded for $x > 1$, we find the upper bound:

$$\log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) \leq n_{ab} \tau \left(\frac{O_{ab}}{n_{ab}} \right) + O(\log n_{ab}).$$

Upper bound, case 2: $n_{ab} = 1$ *and* $O_{ab} = 0$ *or* $n_{ab} = O_{ab}$*, or* $n_{ab} = 0$
In both cases, the corresponding term of the likelihood modularity vanishes, whereas the contribution of the Bayesian modularity is either $\log B(1 + \beta_1, \beta_2)$, $\log(\beta_1, 1 + \beta_2)$, or $\log B(\beta_1, \beta_2)$.

Upper bound, case 3: $n_{ab} \geq 2$ *and* $O_{ab} = 0$ *or* $n_{ab} = O_{ab}$
Again, the corresponding term of the likelihood modularity vanishes. We show the computations for the case $n_{ab} = O_{ab}$; for the case $O_{ab} = 0$, switch β_1 and β_2. By (4.9):

$$\begin{aligned} \log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) &= \log B(n_{ab} + \beta_1, \beta_2) \leq \log \frac{\Gamma(n_{ab} + \lfloor \beta_1 \rfloor + 1)\Gamma(\beta_2)}{\Gamma(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor)} \\ &= (n_{ab} + \lfloor \beta_1 \rfloor) \log \left(\frac{n_{ab} + \lfloor \beta_1 \rfloor}{n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor} \right) + (1/2) \log(n_{ab} + \lfloor \beta_1 \rfloor) \\ &\quad - (\lfloor \beta_1 + \beta_2 \rfloor + 1/2) \log(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor) + \log \Gamma(\beta_2) + \lfloor \beta_1 + \beta_2 \rfloor - 1 + \delta_{ab} - \epsilon_{ab}, \end{aligned}$$

where δ_{ab} and ϵ_{ab} are bounded by constants. Arguing as before, the first term is bounded, while the remainder is of order $\log(n_{ab})$. A lower bound is found analogously.

Lower bound The computations for the lower bound are completely analogous, except that we require $O_{ab} + \beta_1 \geq 2$ and $n_{ab} - O_{ab} + \beta_2 \geq 2$. We study four cases. The cases (1) $O_{ab} \geq 2$ and $n_{ab} - O_{ab} \geq 2$, (2) $n_{ab} = 0$ and (3) $n_{ab} > 0$ and $n_{ab} = O_{ab}$ or $O_{ab} = 0$ are similar to cases 1, 2 and 3 respectively of the upper bound. The fourth case is $n_{ab} - O_{ab} = 1$ and $O_{ab} \geq 2$, or $O_{ab} = 1$ and $n_{ab} - O_{ab} \geq 1$. In both instances, the likelihood modularity is equality to a bounded term minus $\log n_{ab}$. By similar calculations as before, the Bayesian modularity is of the order $\log n_{ab}$ as well.

Conclusion We find:

$$\sum_{a \leq b} \log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) = \sum_{a \leq b} n_{ab} \tau \left(\frac{O_{ab}}{n_{ab}} \right) + O(\log n).$$

Second sum of (4.8)**.**
We consider three cases. If $n_a + \lfloor \alpha \rfloor = 0$, then $\alpha > 0$, implies $n_a = 0$, in which case $\log \Gamma(n_a + \alpha) = \log \Gamma(\alpha)$, which is bounded. In case $n_a + \lfloor \alpha \rfloor = 1$, the term $\log \Gamma(n_a + \alpha)$ is equal to either $\log \Gamma(1 + \alpha)$ or $\log \Gamma(\alpha)$ and thus bounded as well. For the case $n_a + \lfloor \alpha \rfloor \geq 2$, we study the upper bound $\Gamma(n_a + \alpha) \leq \Gamma(n_a + \lfloor \alpha \rfloor + 1)$ and the lower bound $\Gamma(n_a + \alpha) \geq \Gamma(n_a + \lfloor \alpha \rfloor)$. By applying (4.9) in both cases, we conclude:

$$\sum_a \log \Gamma(n_a + \alpha) = \sum_{a : n_a + \lfloor \alpha \rfloor \geq 2} n_a \log n_a - n + O(\log n).$$

□

Lemma 4.8. *For any probability matrix R,*

$$H_P(R) \leq H_P(\mathrm{Diag}(R^T \mathbf{1})\Big). \tag{4.10}$$

Furthermore, if (P, π) is identifiable and the columns of R corresponding to positive coordinates of π are not identically zero, then the inequality is strict unless $P_\sigma R$ is a diagonal matrix for some permutation matrix P_σ.

Proof. This Lemma is related to the proof that the likelihood modularity is consistent given in Bickel and Chen (2009). This proof however rests on their incorrect Lemma 3.1, and thus we provide full details on how the argument can be adapted to avoid the use of their Lemma 3.1 altogether.

For R a diagonal matrix the numbers $(RPR^T)_{ab}/(R\mathbf{1})_a(R\mathbf{1})_b$ reduce to P_{ab}. Consequently, by the definition of H_P,

$$H_P\Big(\mathrm{Diag}(f)\Big) = \sum_{a,b} f_a f_b \, \tau(P_{ab}). \tag{4.11}$$

For a general matrix R, by inserting the definition of τ,

$$\begin{aligned} H_P(R) = &\sum_{a,b} (RPR^T)_{ab} \log \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b} \\ &+ \sum_{a,b} \Big((R\mathbf{1})_a (R\mathbf{1})_b - (RPR^T)_{ab} \Big) \log\Big(1 - \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b}\Big). \end{aligned}$$

Because $(R\mathbf{1})_a(R\mathbf{1})_b - (RPR^T)_{ab} = (R(1-P)R^T)_{ab}$, with 1 the $(K \times K)$-matrix with all coordinates equal to 1, we can rewrite this as

$$\sum_{a,b} \sum_{a',b'} R_{aa'} R_{bb'} \left[P_{a'b'} \log \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b} + (1 - P_{a'b'}) \log\Big(1 - \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b}\Big) \right].$$

By the information inequality for two-point measures, the expressions in square brackets becomes bigger when $(RPR^T)_{ab}/(R\mathbf{1})_a(R\mathbf{1})_b$ is replaced by $P_{a'b'}$, with a strict increase unless these two numbers are equal. After making this substitution the terms in square brackets becomes $\tau(P_{a'b'})$, and we can exchange the order of the two (double) sums and perform the sum on (a, b) to write the resulting expression as

$$\sum_{a',b'} (R^T\mathbf{1})_{a'} (R^T\mathbf{1})_{b'} \tau(P_{a'b'}) = H_P\Big(\mathrm{Diag}(R^T\mathbf{1})\Big).$$

This proves the first assertion (4.10) of the lemma.

If R attains equality, then also for every permutation matrix P_σ, by the equality $H_P(P_\sigma R) = H_P(R)$ and the fact that $(P_\sigma R)^T\mathbf{1} = R^T\mathbf{1}$, we have

$$H_P(P_\sigma R) = H_P\Big(\mathrm{Diag}((P_\sigma R)^T\mathbf{1})\Big). \tag{4.12}$$

We shall show that if R satisfies this equality and $P_\sigma R$ has a positive diagonal, then $P_\sigma R$ is in fact diagonal. Furthermore, we shall show that there exists P_σ such that $P_\sigma R$ has a positive diagonal.

Fix some $(P_\sigma)_m$ that maximizes the number of positive diagonal elements of $P_\sigma R$ over all permutation matrices P_σ, and denote $\bar{R} = (P_\sigma)_m R$. Because the information inequality is strict, the preceding argument shows that (4.12) can be true for $P_\sigma = (P_\sigma)_m$ (giving $P_\sigma R = \bar{R}$) only if

$$P_{a'b'} = \frac{(\bar{R}P\bar{R}^T)_{ab}}{(\bar{R}\mathbf{1})_a(\bar{R}\mathbf{1})_b}, \qquad \text{whenever } \bar{R}_{aa'}\bar{R}_{bb'} > 0. \tag{4.13}$$

Denote the matrix on the right of the equality by Q.

If $\bar{R}$ has a completely positive diagonal, then we can choose $a = a'$ and $b = b'$ and find from equation (4.13), that $P_{ab} = Q_{ab}$, for every a, b. If also $\bar{R}_{aa'} > 0$, then we can also choose $b = b'$ and find that $P_{a'b} = Q_{ab}$, for every b. Thus the ath and a'th rows of P are identical. Since all rows of P are different by assumption, it follows that no $a \neq a'$ with $\bar{R}_{aa'} > 0$ exists.

If $\bar{R}$ does not have a fully positive diagonal, then the submatrix of $\bar{R}$ obtained by deleting the rows and columns corresponding to positive diagonal elements must be the zero matrix, since otherwise we might permute the remaining rows and create an additional nonzero diagonal element, contradicting that $(P_\sigma)_m$ already maximized this number. If I and I^c are the sets of indices of zero and nonzero diagonal elements, then the preceding observation is that $\bar{R}_{ij}$ is zero for every $i, j \in I$. If $\pi > 0$, then we need to consider only R with nonzero columns. For $i \in I$ a nonzero element in the ith column of $\bar{R}$ must be located in the rows with label in I^c: for every $i \in I$ there exists $k_i \in I^c$ with $\bar{R}_{k_i i} > 0$. Then, for $i, j \in I$,

(1) for $a = k_i$, $b = k_j$, $a' = i$, $b' = j$, equation (4.13) implies $Q_{k_i k_j} = P_{ij}$.

(2) for $a = k_i$, $b \in I^c$, $a' = i$, $b' = b$, equation (4.13) implies $Q_{k_i b} = P_{ib}$.

(3) for $a = k_i$, $b \in I^c$, $a' = k_i$, $b' = b$, equation (4.13) implies $Q_{k_i b} = P_{k_i b}$.

We combine these three assertions to conclude that, for $a, i \in I$ and $b \in I^c$,

$$\begin{aligned} P_{ai} = P_{ia} &\overset{(1)}{=} Q_{k_i k_a} \overset{(2)}{=} P_{ik_a} = P_{k_a i}, \\ P_{ab} &\overset{(2)}{=} Q_{k_a b} \overset{(3)}{=} P_{k_a b}. \end{aligned}$$

Together these imply that the ath and the k_ath row of P are equal. Since by assumption they are not (if $\pi > 0$), this case can actually not exist (i.e. $k = 0$).

Finally if $\pi_a = 0$ for some a, then we follow the same argument, but we match only every column $i \in I$ with $\pi_i > 0$ to a row $k_i \in I^c$. By the assumption on R such k_i exist, and the construction results in two rows of P that are identical in the coordinates with $\pi_a > 0$. □

Lemma 4.9. *For any fixed $(K \times K)$-matrix P with elements in $[0,1]$, uniformly in probability matrices R, as $\rho_n \to 0$,*

$$\frac{1}{\rho_n}\Big(H_{\rho_n P}(\mathrm{Diag}(R^T \mathbf{1})) - H_{\rho_n P}(R)\Big) \to G_P(\mathrm{Diag}(R^T \mathbf{1})) - G_P(R). \tag{4.14}$$

Furthermore, if (P,π) is identifiable and the columns of R corresponding to positive coordinates of π are not identically zero, then the right side is strictly positive unless SR is a diagonal matrix for some permutation matrix S.

Proof. From the fact that $|(1-u)\log(1-u)+u| \le u^2$, for $0 \le u \le 1$, it can be verified that, $\left|\rho_n^{-1}\tau(\rho_n u) - \Big(u\log\rho_n + \tau_0(u)\Big)\right| \le \rho_n \to 0$, uniformly in $0 \le u \le 1$. It follows that, uniformly in R,

$$\frac{1}{\rho_n}H_{\rho_n P}(R) = \log\rho_n \sum_{a,b}(RPR^T)_{ab} + \sum_{a,b}(R\mathbf{1})_a(R\mathbf{1})_b\tau_0\Big(\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\Big) + O(\rho_n).$$

The first term on the right is equal to $\log\rho_n(R^T\mathbf{1})^T P(R^T\mathbf{1})$, and hence is the same for R and $\mathrm{Diag}(R^T\mathbf{1})$. Thus this term cancels on taking the difference to form the left side of (4.14), and hence (4.14) follows.

The right side of (4.14) is nonnegative, because the left side is, by Lemma 4.8. This fact can also be proved directly along the lines of the proof of Lemma 4.8, as follows. Write

$$G_P(R) = \sum_{a,b}\sum_{a',b'} R_{aa'}R_{bb'}\Big[P_{a'b'}\log\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b} - \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\Big].$$

By the information inequality for two Poisson distributions the term in square brackets becomes bigger if $(RPR^T)_{ab}/(R\mathbf{1})_a(R\mathbf{1})_b$ is replaced by $P_{a'b'}$. It then becomes $\tau_0(P_{a'b'})$ and the double sum on (a,b) can be executed to see that the resulting bound is $G_P\Big(\mathrm{Diag}(R^T\mathbf{1})\Big)$. Furthermore, the inequality is strictly unless (4.13) holds, with $\bar{R} = R$. Since also $G_P(P_\sigma R) = G_P(R)$, for every permutation matrix P_σ, the final assertion of the lemma is proved by copying the proof of Lemma 4.8. □

4.6.2 Strong consistency

We need slightly adapted versions of the function H_P, given by, with δ_{ab} equal to 1 or 0 if $a = b$ or not,

$$H_{P,n}(R) = \frac{1}{2}\sum_{a,b}(R\mathbf{1})_a\Big((R\mathbf{1})_b - \delta_{ab}/n\Big)\,\tau\Big(\frac{(RPR^T)_{ab} - \delta_{ab}\sum_k P_{kk}R_{ka}/n}{(R\mathbf{1})_a\Big((R\mathbf{1})_b - \delta_{ab}/n\Big)}\Big). \tag{4.15}$$

For given functions $t_{ab} : [0,1] \to \mathbb{R}$, let $X(e)$ be the $K \times K$ matrix with entries

$$X_{ab}(e) = t_{ab}\Big(\frac{\widetilde{O}_{ab}(e)}{n^2}\Big) - t_{ab}\Big(\frac{\mathbb{E}(\widetilde{O}_{ab}(e) \mid Z)}{n^2}\Big). \tag{4.16}$$

Proof of Theorem 4.3 [strong consistency]

Proof. i. By Theorem 4.4, $\widehat{e}$ is weakly consistent, and hence with probability tending to one it belongs to the set of classifications e such that the fractions $f(e)$ are close to π, and the matrices $R(e,Z)$ are close to $\mathrm{Diag}(\pi)$ after the appropriate permutation of the labels (that is, of rows of $R(e,Z)$). Therefore, it is no loss of generality to assume that $\widehat{e}$ is restricted to this set. By Lemmas 4.5 and 4.6, the matrices $\widetilde{O}(e)/n^2$ are then close to

$R(e,Z)PR(e,Z)^T \to \mathrm{Diag}(\pi)P\mathrm{Diag}(\pi)$, and hence are bounded away from zero and one if P has this property.

If $\widehat{e}$ and Z differ at m nodes, then $\widehat{e}$ belongs to the set of e with $\|R(Z,Z) - R(e,Z)\|_1 = m(2/n)$, by Lemma 4.1. In that case $Q_B(e) \geq Q_B(Z)$, for some e in this set, and hence by Lemma 4.2 $Q_{ML}(e) - Q_{ML}(Z) + Q_P(e) - Q_P(Z) \geq -\eta_n$, for some η_n of order $(\log n)/n^2$. It follows that:

$$\begin{aligned}\Big[Q_{ML}(e) - H_{P,n}\Big(R(e,Z)\Big)\Big] - \Big[Q_{ML}(Z) - H_{P,n}\Big(R(Z,Z)\Big)\Big] \\ \geq H_{P,n}\Big(R(Z,Z)\Big) - H_{P,n}\Big(R(e,Z)\Big) - |Q_P(e) - Q_P(Z)| - \eta_n. \end{aligned} \tag{4.17}$$

The first term on the right is bounded below by a multiple of m/n, by Lemmas 4.10 and 4.1. Because $(x+\alpha)\log x - (y+\alpha)\log y = \int_x^y (\log s + (s+\alpha)/s)\,ds$ is bounded in absolute value by a multiple of $|x-y|\log(x \vee y)$, if $\alpha \geq 0$ and $x, y > 0$, the second term $-|Q_P(e) - Q_P(Z)|$ is bounded below by a multiple of $m(\log n)/n^2$, for some positive constant C_2, which is of smaller order than m/n. We conclude that the left side of (4.17) is bounded below by $C_1 m/n$. The left side is $\sum_{a,b}\Big(X_{ab}(e) - X_{ab}(Z)\Big)$, for X defined in (4.16) and t the function with coordinates $t_{ab}(o) = f_a(e)\Big(f_b(e) - \delta_{ab}/n\Big)\tau\Big(o/f_a(e)\Big(f_b(e) - \delta_{ab}/n\Big)\Big)$. Because we restrict e to classifications such that $O_{ab}(e)/n_{ab}(e)$ and $f_a(e)f_b(e)$ are bounded away from zero and one, only the values of the function τ on an open interval strictly within $(0,1)$ matter. On any such interval τ has uniformly bounded derivatives, and hence the bound of Lemma 4.13 is valid. Thus we find that

$$\begin{aligned}\Pr\Big(\#(i : \widehat{e}_i \neq Z_i) = m\Big) &\leq \Pr\Big(\sup_{e:\#(i:e_i \neq Z_i)\leq m} \Big\|X(e) - X(Z)\Big\|_\infty \geq \frac{C_1 m}{n}\Big) \\ &\lesssim K^m \binom{n}{m} e^{-cm^2/(m\|P\|_\infty/n + m/n)} \\ &\leq e^{m\log(Kne/m) - c_1 mn}. \end{aligned}$$

The sum of the right side over $m = 1, \ldots, n$ tends to zero.

ii. We follow the proof for i, but in (4.17) use that $H_{P,n}\Big(R(Z,Z)\Big) - H_{P,n}\Big(R(e,Z)\Big) \geq \rho_n C\|R(Z,Z) - R(e,Z)\|_1 \geq \rho_n C 2m/n$, by Lemma 4.12. Since $\rho_n \gg (\log n)/n$ by assumption, we have that the contribution $m(\log n)/n^2$ of $Q_P(e) - Q_P(Z)$ is still negligible and hence $\rho_n C 2m/n$ is a lower bound for the left side of (4.17). As a bound on the left side of the preceding display, we then obtain

$$\sum_{m=1}^n K^m \binom{n}{m} e^{-c_2\rho_n^2 m^2/(m\rho_n/n + \rho_n m/n)} \leq \sum_{m=1}^n e^{m\log(Kne/m) - c_3\rho_n mn}.$$

This sum tends to zero provided that $n\rho_n \gg \log n$. □

Lemma 4.10. *If P is fixed and symmetric and every pair of rows of P is different and $0 < P < 1$ and $\pi > 0$, then, for sufficiently small $\delta > 0$,*

$$\liminf_{n\to\infty} \inf_{0 < \|R - \mathrm{Diag}(\pi)\| < \delta} \frac{H_{P,n}\Big(\mathrm{Diag}(R^T \mathbf{1})\Big) - H_{P,n}(R)}{\|\mathrm{Diag}(R^T \mathbf{1}) - R\|} > 0. \tag{4.18}$$

Proof. We can reparametrize the $K \times K$ matrices R by the pairs $(R^T\mathbf{1}, R - \mathrm{Diag}(R^T\mathbf{1}))$, consisting of the K vector $f = R^T\mathbf{1}$ and the $K \times K$ matrix $R - \mathrm{Diag}(R^T\mathbf{1})$. The latter matrix is characterized by having nonnegative off-diagonal elements and zero column sums, and can be represented in the basis consisting of all $K \times K$ matrices $\Delta_{bb'}$, for $b \neq b'$, defined by: $(\Delta_{bb'})_{b'b'} = -1$, $(\Delta_{bb'})_{bb'} = 1$ and $(\Delta_{bb'})_{aa'} = 0$, for all other entries (a, a'), i.e. the b'th column of $\Delta_{bb'}$ has a 1 in the bth coordinate and a -1 on the b'th coordinate and all its other columns are zero. Given any matrix $R \geq 0$ the matrix $R - \mathrm{Diag}(R^T\mathbf{1})$ can be decomposed as

$$R - \mathrm{Diag}(R^T\mathbf{1}) = \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'},$$

for $\lambda_{bb'} = R_{bb'} \geq 0$. Since every $\Delta_{bb'}$ has exactly one nonzero off-diagonal element, which is equal to 1, and in a different location for each $b \neq b$, the sum of the off-diagonal elements of the matrix on the right side is $\sum_{b,b'} \lambda_{bb'}$. Because the sum of all its elements is zero, it follows that its sum of absolute elements is given by $\|R - \mathrm{Diag}(R^T\mathbf{1})\|_1 = 2\sum_{b \neq b'} \lambda_{bb'}$.

Thus we obtain a further reparametrization $R \leftrightarrow (f, \lambda)$, in which $R = \mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}$. For given P, f and n, define the function

$$G(\lambda) = H_{P,n}\Big(\mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}\Big).$$

Then we would like to show that there exists C such that

$$\frac{H_{P,n}(\mathrm{Diag}(R^T\mathbf{1})) - H_{P,n}(R)}{\|R - \mathrm{Diag}(R^T\mathbf{1})\|_1} = \frac{G(0) - G(\lambda)}{2\sum_{b \neq b'} \lambda_{bb'}} \geq C > 0,$$

for every f in a neighbourhood of π, λ in a neighbourhood of 0 intersected with $\{\lambda : \lambda \geq 0\}$, and every sufficiently large n. The numerator in the quotient is $f(0) - f(1)$ for the function $f(s) = G(s\lambda)$. Writing this difference in the form $-f'(0) - \int_0^1 \big(f'(s) - f'(0)\big)\, ds$ gives that the numerator is equal to

$$-\nabla G(0)^T \lambda - \int_0^1 \Big(\nabla G(s\lambda) - \nabla G(0)\Big)\Big)^T\, ds\, \lambda. \tag{4.19}$$

It suffices to show that the first term is bounded below by a multiple of $\|\lambda\|_1$ and that the second is negligible relative to the first, as $n \to \infty$, uniformly in f in a neighbourhood of π and λ in a neighbourhood of 0 intersected with $\{\lambda : \lambda \geq 0\}$. Thus it is sufficient to show first that for every coordinate $\lambda_{bb'}$ of λ minus the partial derivative of G at $\lambda = 0$ with respect to $\lambda_{bb'}$ is bounded away from 0, as $n \to \infty$ uniformly in f, and second that every partial derivative is equicontinuous at $\lambda = 0$ uniformly in f and large n.

We have

$$G(\lambda) = \frac{1}{2} \sum_{a,a'} f_a(\lambda)\Big(f_{a'}(\lambda) - \delta_{aa'}/n\Big)\, \tau\Big(\frac{\Big(R(\lambda)PR(\lambda)^T\Big)_{aa'} - \delta_{aa'} e_a(\lambda)/n}{f_a(\lambda)\Big(f_{a'}(\lambda) - \delta_{aa'}/n\Big)}\Big), \tag{4.20}$$

for

$$f(\lambda) = f + \sum_{bb'} \lambda_{bb'}(\Delta_{bb'}\mathbf{1}),$$

$$R(\lambda) = \mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'},$$
$$e_a(\lambda) = \sum_k P_{kk} R_{ak}(\lambda) = P_{aa} f_a + \sum_{b \neq b'} P_{b'b'} \lambda_{bb'} (\delta_{ab} - \delta_{ab'}).$$

By a lengthy calculation, given in Lemma 4.11,

$$\frac{\partial}{\partial \lambda_{bb'}} G(\lambda)_{|\lambda=0} = -\sum_a f_a K(P_{ab'} \| P_{ab}) + \frac{1}{2n} K(P_{b'b'} \| P_{bb}), \tag{4.21}$$

for $K(p\|q) = p \log(p/q) + (1-p)\log\Big((1-p)/(1-q)\Big)$ the Kullback-Leibler divergence between the Bernoulli distributions with success probabilities p and q. The numbers f_a are bounded away from zero for f sufficiently close to π, and hence so is $\sum_a f_a K(P_{ab'} \| P_{ab})$, unless the bth and b'th column of P are identical. The whole expression is bounded below by the minimum over (b, b') of these numbers minus $(2n)^{-1}$ times the maximum of the numbers $K(P_{b'b'} \| P_{bb})$, and hence is positive and bounded away from zero for sufficiently large n.

To verify the equicontinuity of the partial derivatives we can compute these explicitly at λ and take their limit as $n \to \infty$. We omit the details of this calculation. However, we note that every term of $G(\lambda)$ is a fixed function of the quadratic forms in λ

$$\Big(f_a + \sum_{bb'} \lambda_{bb'} (\Delta_{bb'} \mathbf{1})_a\Big)\Big(f_{a'} + \sum_{bb'} \lambda_{bb'} (\Delta_{bb'} \mathbf{1})_{a'} - \delta_{aa'}/n\Big), \tag{4.22}$$
$$\Big(\Big(\mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}\Big) P \Big(\mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}^T\Big)\Big)_{aa'}$$
$$- \frac{\delta_{aa'}}{2n}\Big(P_{aa} f_a + \sum_{b \neq b'} P_{b'b'} \lambda_{bb'} (\delta_{ab} - \delta_{ab'})\Big). \tag{4.23}$$

These forms are obviously smooth in λ, and their dependence and that of their derivatives on n is seen to vanish as $n \to \infty$. For f and λ restricted to neighbourhoods of π and 0, the values of the quadratic forms are restricted to a domain in which the transformation mapping them into $G(\lambda)$ is continuously differentiable. Thus the desired equicontinuity follows by the chain rule. □

Lemma 4.11. *The partial derivatives of the function G at 0 defined by* (4.20) *are given by* (4.21).

Proof. For given differentiable functions u and v the map $\epsilon \mapsto u(\epsilon)\tau\Big(v(\epsilon)/u(\epsilon)\Big)$ has derivative $v' \log\Big(v/(u-v)\Big) - u' \log\Big(u/(u-v)\Big)$. We apply this for every given pair (a, a') to the functions u and v obtained by taking $\lambda_{bb'}$ in (4.22) and (4.23) equal to ϵ and all other coordinates of λ equal to zero. Then

$$u(0) = f_a(f_{a'} - \delta_{aa'}/n),$$
$$v(0) = f_a(f_{a'} - \delta_{aa'}/n) P_{aa'},$$
$$u'(0) = (\Delta_{bb'} \mathbf{1})_a (f_{a'} - \delta_{aa'}/n) + f_a (\Delta_{bb'} \mathbf{1})_{a'}$$

$$v'(0) = (\Delta_{bb'}P)_{aa'}f_{a'} + f_a(\Delta_{bb'}P)_{a'a} - (\delta_{aa'}/n)P_{b'b'}(\delta_{ab} - \delta_{ab'}).$$

It follows that $v(0)/(u(0) - v(0)) = P_{aa'}/(1 - P_{aa'})$, and $u(0)/(u(0) - v(0)) = 1/(1 - P_{aa'})$. Hence in view of (4.15) the partial derivative in (4.21) is equal to

$$\sum_{a \neq a'} \left[v'(0) \log \frac{P_{aa'}}{1 - P_{aa'}} - u'(0) \log \frac{1}{1 - P_{aa'}} \right].$$

We combine this with the equalities

$$(\Delta_{bb'}\mathbf{1})_a = \begin{cases} 0 & \text{if } a \notin \{b, b'\}, \\ -1 & \text{if } a = b', \\ 1 & \text{if } a = b, \end{cases} \qquad (\Delta_{bb'}P)_{aa'} = \begin{cases} 0 & \text{if } a \notin \{b, b'\}, \\ -P_{b'a'} & \text{if } a = b', \\ P_{b'a'} & \text{if } a = b. \end{cases}$$

□

Lemma 4.12. *If S is fixed and symmetric, every pair of rows of S is different and $S > 0$ and $\pi > 0$ coordinatewise, then there exists $C > 0$ such that, for sufficiently small $\delta > 0$ and any $\rho_n \downarrow 0$,*

$$\liminf_{n \to \infty} \inf_{0 < \|R - \mathrm{Diag}(\pi)\| < \delta} \frac{H_{\rho_n S, n}\left(\mathrm{Diag}(R^T \mathbf{1})\right) - H_{\rho_n S, n}(R)}{\rho_n \|\mathrm{Diag}(R^T \mathbf{1}) - R\|} \geq C.$$

Proof. In the notation of the proof of Lemma 4.10 we must now show that $G(0) - G(\lambda) \geq C\rho_n \|\lambda\|_1$, as $n \to \infty$, uniformly in f in a neighbourhood of π, and λ in a positive neighbourhood of 0. As in that proof we write $G(0) - G(\lambda)$ in the form (4.19) and see that it suffices that the partial derivatives of G at 0 divided by ρ_n tend to negative limits, and that $\left\| \nabla G(\lambda) - \nabla G(0) \right\| / \rho_n$ becomes uniformly small as λ is close enough to zero.

The partial derivative at 0 with respect to $\lambda_{bb'}$ is given in (4.21), where we must replace P by $\rho_n S$. Since the scaled Kullback-Leibler divergence $\rho_n^{-1} K(\rho_n s \| \rho_n t)$ of two Bernoulli laws converges to the Kullback-Leibler divergence $K_0(s\|t) = s \log(s/t) + t - s$ between two Poisson laws of means s and t, as $\rho_n \to 0$, it follows that for $\rho_n \to 0$, uniformly in f,

$$\frac{1}{\rho_n} \frac{\partial}{\partial \lambda_{bb'}} G(\lambda)_{|\lambda = 0} \to - \sum_a f_a K_0(S_{ab'} \| S_{ab}).$$

The right side is strictly negative by the assumption that every pair of rows of S differ in at least one coordinate.

If $P = \rho_n S$, then the function $\lambda \mapsto v(\lambda)$ given in (4.23) takes the form $v = \rho_n v_S$, for v_S defined in the same way but with S replacing P. The function u given in (4.22) does not depend on P or S. Using again that the derivative of the map $\epsilon \mapsto u(\epsilon)\tau\left(v(\epsilon)/u(\epsilon)\right)$ is given by $v' \log\left(v/(u - v)\right) - u' \log\left(u/(u - v)\right)$, we see that the partial derivative with respect to $\lambda_{bb'}$ of the (a, a') term in the sum defining G takes the form

$$\begin{aligned} &\rho_n v_S' \log \frac{\rho_n v_S}{u - \rho v_S} - u' \log \frac{u}{u - \rho_n v_S} \\ &\quad = \rho_n v_S' \log \rho_n - \rho_n v_S' \log(v_S/u) - (\rho_n v_S' - u') \log(1 - \rho_n v_S/u). \end{aligned}$$

Here u and V_S are as in (4.22) and (4.23) (with P replaced by S), and depend on (a, a'). From the fact that the column sums of the matrices $R(\lambda)$ do not depend on λ, we have that

$$\sum_{a,a'}\Big[(R(\lambda)SR(\lambda)^T)_{aa'} - \frac{\delta_{aa'}}{n}\sum_k P_{kk}R(\lambda)_{ak}\Big] = R(\lambda)^T\mathbf{1}SR(\lambda)^T\mathbf{1} - \sum_k P_{kk}\sum_a R(\lambda)_{ak}$$

is constant in λ. This shows that $\sum_{a,a'} v'_S = 0$ and hence the contribution of the term $\rho_n v'_S \log\rho_n$ to the partial derivatives of G vanishes. The term $-(\rho_n v'_S - u')\log(1-\rho_n v_S/u)$ can be expanded as $(\rho_n v'_S - u')\rho_n v_S/u$ up to $O(\rho_n^2)$, uniformly in f and λ. Since these are equicontinuous functions of λ, it follows that $\rho_n^{-1}\big(\nabla G(\lambda) - \nabla G(0)\big)$ becomes arbitrarily small if λ varies in a sufficiently small neighbourhood of 0. □

Lemma 4.13. *There exists a constant $c > 0$ such that for $X(e)$ as in* (4.16)*, for every twice differentiable functions $t_{a,b} : [0,1] \to \mathbb{R}$ with $\|t'_{a,b}\|_\infty \vee \|t''_{a,b}\|_\infty \le 1$, and every $x > 0$,*

$$\Pr\Big(\max_{e:\#(e_i\neq Z_i)\le m}\big\|X(e) - X(Z)\big\|_\infty > x\Big) \qquad \le 6\binom{n}{m}K^{m+2}e^{-\frac{cx^2n^2}{m\|P\|_\infty/n+x}}.$$

Proof. Given Z there are at most $\binom{n}{m}$ groups of m candidate nodes that can be assigned to have $e_i \neq Z_i$, and the label of each node can be chosen in at most $K-1$ ways. Thus conditioning the probability on Z, we can use the union bound to pull out the maximum over e, giving a sum of fewer than $\binom{n}{m}K^m$ terms. Next we pull out the norm giving another factor K^2. It suffices to combine this with a tail bound for a single variable $X_{a,b}(e) - X_{a,b}(Z)$. Write t for $t_{a,b}$.

Assume for simplicity of notation that $e_i = Z_i$, for $i > m$, and decompose

$$\begin{aligned}\frac{1}{n^2}O_{ab}(e) &= \frac{1}{n^2}\Big[\sum_{i\le m \text{ or } j\le m} A_{ij}1_{e_i=a,e_j=b} + \sum_{i>m \text{ and } j>m} A_{ij}1_{e_i=a,e_j=b}\Big]\\ &=: S_1 + S_2.\end{aligned}$$

Let $O_{ab}(Z)/n^2 =: S'_1 + S_2$, with the same variable S_2, be the corresponding decomposition if e is changed to Z, and then decompose, where the expectation signs $\mathbb{E}$ denote conditional expectations given Z,

$$\begin{aligned}&X_{ab}(e) - X_{ab}(Z)\\ &\quad= \Big(t(S_1+S_2) - t(\mathbb{E}S_1 + \mathbb{E}S_2)\Big) - \Big(t(S'_1+S_2) - t(\mathbb{E}S'_1 + \mathbb{E}S_2)\Big)\\ &\quad= t(S_1+S_2) - t(\mathbb{E}S_1 + S_2)\\ &\qquad+ \Big(t(\mathbb{E}S_1+S_2) - t(\mathbb{E}S_1 + \mathbb{E}S_2)\Big) - \Big(t(\mathbb{E}S'_1+S_2) - t(\mathbb{E}S'_1 + \mathbb{E}S_2)\Big)\\ &\qquad+ t(\mathbb{E}S'_1+S_2) - t(S'_1+S_2)\end{aligned}$$

The first and third terms on the far right can be bounded above in absolute value by $\|t'\|_\infty$ times the increment. To estimate the second term we write it as

$$(S_2 - \mathbb{E}S_2)(\mathbb{E}S_1 - \mathbb{E}S'_1)\int_0^1\int_0^1 t''\Big(uS_2 + (1-u)\mathbb{E}S_2 + v\mathbb{E}S_1 + (1-v)\mathbb{E}S'_1\Big)\,du\,dv.$$

Since the first and second derivatives of t are uniformly bounded by 1, it follows that

$$\left|X_{ab}(e) - X_{ab}(Z)\right| \leq |S_1 - \mathbb{E}S_1| + |S_2 - \mathbb{E}S_2|\,|\mathbb{E}S_1 - \mathbb{E}S_1'| + |S_1' - \mathbb{E}S_1'|.$$

The variable $S_1 - \mathbb{E}S_1$ is a sum of fewer than $2mn$ independent variables, each with conditional mean zero, bounded above by $1/n^2$ and of variance bounded above by $\|P\|_\infty/n^4$. Therefore Bernstein's inequality gives that

$$\mathbb{P}\Big(|S_1 - \mathbb{E}S_1| > x\Big) \leq e^{-\frac{1}{2}x^2/(2mn\|P\|_\infty/n^4 + x/(3n^2))}.$$

This is as the exponential factor in the bound given by the lemma, for appropriate c. The variable $S_1' - \mathbb{E}S_1'$ can be bounded similarly. Furthermore $|\mathbb{E}S_1 - \mathbb{E}S_1'| \leq 4mn/n^2 = 4m/n$, and $S_2 - \mathbb{E}S_2$ is the sum of fewer than n^2 variables as before, so that

$$\mathbb{P}\Big(|S_2 - \mathbb{E}S_2|\,|\mathbb{E}S_1 - \mathbb{E}S_1'| > x\Big) \leq e^{-\frac{1}{2}(xn/(4m))^2/(n^2\|P\|_\infty/n^4 + xn/(12mn^2))}.$$

The exponent has a similar form as before, except for an additional factor $n/m \geq 1$. □

5

The switch criterion in nested model selection

Abstract

We study the switch distribution, introduced by Van Erven et al. (2012), applied to model selection and subsequent estimation. While switching was known to be strongly consistent, here we show that it achieves minimax optimal parametric risk rates up to a $\log\log n$ *factor when comparing two nested exponential families, partially confirming a conjecture by Lauritzen (2012) and Cavanaugh (2012) that switching behaves asymptotically like the Hannan-Quinn criterion. Moreover, like Bayes factor model selection but unlike standard significance testing, when one of the models represents a simple hypothesis, the switch criterion defines a* robust *null hypothesis test, meaning that its Type-I error probability can be bounded irrespective of the stopping rule. Hence, switching is consistent, insensitive to optional stopping and almost minimax risk optimal, showing that, Yang's (2005) impossibility result notwithstanding, it is possible to 'almost' combine the strengths of AIC and Bayes factor model selection.*

5.1 Introduction

We consider the following standard model selection problem, where we have i.i.d. observations $X_1, \ldots, X_n$ and we wish to select between two nested parametric models,

$$\mathcal{M}_0 = \left\{p_\mu \mid \mu \in M_0\right\} \quad \text{and} \quad \mathcal{M}_1 = \{p_\mu \mid \mu \in M_1\}. \tag{5.1}$$

This chapter is set to appear in *Statistica Sinica*, as: S. van der Pas and P. Grünwald. Almost the best of three worlds: risk, consistency and optional stopping for the switch criterion in nested model selection. The central result of this paper, Theorem 5.5, already appeared in the Master's Thesis (Van der Pas, 2013) for the special case where $m_1 = 1$ and $m_0 = 0$. This research was supported by NWO VICI Project 639.073.04.

Here the X_i are random vectors taking values in some set $\mathcal{X}$, $M_1 \subseteq \mathbb{R}^{m_1}$ for some $m_1 > 0$ and $\mathcal{M}_0 = \{p_\mu : \mu \in M_0\} \subset \mathcal{M}_1$ represents an m_0-dimensional submodel of $\mathcal{M}_1$, where $0 \leq m_0 < m_1$. We may thus denote $\mathcal{M}_0$ as the 'simple' and $\mathcal{M}_1$ as the 'complex' model. We will assume that $\mathcal{M}_1$ is an exponential family, represented as a set of densities on $\mathcal{X}$ with respect to some fixed underlying measure, so that p_μ represents the density of the observations, and we take it to be given in its mean-value parameterization. As the notation indicates, we require, without loss of generality, that the parameterizations of $\mathcal{M}_0$ and $\mathcal{M}_1$ coincide, that is $M_0 \subset M_1$ is itself a set of m_1-dimensional vectors, the final $m_1 - m_0$ components of which are fixed to known values. We restrict ourselves to the case in which both M_1 and the restriction of M_0 to its first m_0 components are products of open intervals. Most model selection methods output not just a decision $\delta(X^n) \in \{0,1\}$, but also an indication $r(X^n) \in \mathbb{R}$ of the strength of evidence, such as a p-value or a Bayes factor. As a result, such procedures can often be interpreted as methods for hypothesis testing, where $\mathcal{M}_0$ represents the *null* model and $\mathcal{M}_1$ the alternative; a very simple example of our setting is when the X_i consist of two components $X_i \equiv (X_{i1}, X_{i2})$, which according to $\mathcal{M}_1$ are independent Gaussians whereas under $\mathcal{M}_2$ they can have an arbitrary bivariate Gaussian distribution and hence can be dependent. Since we allow $\mathcal{M}_0$ to be a singleton, this setting also includes some very simple, classical yet important settings such as testing whether a coin is biased ($\mathcal{M}_0$ is the fair coin model, $\mathcal{M}_1$ contains all Bernoulli distributions).

We consider three desirable properties of model selection methods: (a) optimal worst-case risk rate of post-model selection estimation (with risk measured in terms of squared error loss, squared Hellinger distance, Rényi or Kullback-Leibler divergence); (b) consistency, and, (c) for procedures which also output a strength of evidence $r(X^n)$, whether the validity of the evidence is insensitive to optional stopping under the null model. We evaluate the recently introduced model selection criterion δ_{sw} based on the switch distribution (Van Erven et al., 2012) on properties (a), (b) and (c).

The switch distribution, introduced by Van Erven et al., (2007), was originally designed to address the *catch-up phenomenon*, which occurs when the best predicting model is not the same across sample sizes. The switch distribution can be interpreted as a modification of the Bayesian predictive distribution. It also has an MDL interpretation: if one corrects standard MDL approaches (Grünwald, 2007) to take into account that the best predicting method changes over time, one naturally arrives at the switch distribution. Lhéritier and Cazals (2015) describes a successful practical application for two-sample sequential testing, related to the developments in this paper but in a nonparametric context. We briefly give the definitions relevant to our setting in Section 5.2; for all further details we refer to Van Erven et al. (2012) and 5.7.5 in Section 5.7.

When evaluating any model selection method, there is a well-known tension between (a) and (b): the popular AIC method (Akaike, 1973) achieves the minimax optimal parametric rate of order $1/n$ in the problem above, but is inconsistent; the same holds for the many popular model selection methods that asymptotically tend to behave like AIC, such as k-fold and leave-one-out-cross-validation, the bootstrap and Mallow's C_p in linear regression (Efron, 1986; Shao, 1997; Stone, 1977). On the other hand, BIC (Schwarz, 1978) is consistent in the sense that for large enough n, it will select the smallest model containing the 'true' μ; but it misses the minimax parametric rate by a factor of $\log n$. The same holds for traditional Minimum Description Length (MDL) approaches (Grünwald, 2007)

and Bayes factor model selection (BFMS) (Kass and Raftery, 1995), of which BIC is an approximation. This might lead one to wonder if there exists a single method that is optimal in both respects. A key result by Yang (2005) shows that this is impossible: any consistent method misses the minimax optimal rate by a factor $g(n)$ with $\lim_{n\to\infty} g(n) = \infty$.

In Section 5.4.2 we show that, Yang's result notwithstanding, the switch distribution allows us to get very close to satisfying property (a) and (b) at the same time, at least in the problem defined above (Yang's result was shown in a nested linear regression rather than our exponential family context, but it does hold in our exponential family setting as well; see the discussion at the end of Section 5.3.3). We prove that in our setting, the switch model selection criterion δ_{sw} (a) misses the minimax optimal rate only by an exceedingly small $g_{\text{sw}}(n) \asymp \log\log n$ factor (Theorem 5.5). Property (b), strong consistency, was already shown by Van Erven et al. (2012). The factor $g_{\text{sw}}(n) \asymp \log\log n$ is an improvement over the extra factor resulting from Bayes factor model selection, which has $g_{\text{BFMS}}(n) \asymp \log n$. Indeed, as discussed in the introduction of Van Erven et al. (2012), the catch-up phenomenon that the switch distribution addresses is intimately related to the rate-suboptimality of Bayesian inference. Van Erven et al. (2012) show that, while model selection based on switching is consistent, sequential prediction based on model averaging with the switching method achieves minimax optimal *cumulative* risk rates in general parametric and nonparametric settings, where the cumulative risk at sample size n is obtained by summing the standard, instantaneous risk from 1 to n. In contrast, in nonparametric settings, standard Bayesian model averaging typically has a cumulative risk rate that is larger by a $\log n$ factor. Using the cumulative risk is natural in sequential prediction settings, but Van Erven et al. (2012) left open the question of how switching would behave for the more standard, instantaneous risk. In contrast to the cumulative setting, we cannot expect to achieve the optimal rate here by Yang's (2005) result, but it is interesting to see that switching gets so close.

We now turn to the third property, robustness to optional stopping. While consistency in the sense above is an asymptotic and even somewhat controversial notion (see Section 5.6), there exists a nonasymptotic property closely related to consistency that, while arguably much more important in practice, has received relatively little attention in the recent statistical literature. This is property (c) above, insensitivity to optional stopping. In statistics, the issue was thoroughly discussed, yet never completely resolved, in the 1960s; nowadays, it is viewed as a highly desirable feature of testing methods by, for example, psychologists; see (Sanborn and Hills, 2014; Wagenmakers, 2007). In particular, it is often argued (Wagenmakers, 2007) that the fixed stopping rule required by the classical Neyman-Pearson paradigm severely and unnecessarily restricts the application domain of hypothesis testing, invalidating much of the p-values reported in the psychological literature. Approximately 55% of psychologists admitted in a survey to deciding whether to collect more data after looking at their results to see if they were significant (John et al., 2012). We analyze property (c) in terms of *robust null hypothesis tests*, formally defined in Section 5.5. A method defines a robust null hypothesis test if (1) it outputs evidence $r(X^n)$ that does not depend on the stopping rule used to determine n, and (2) (some function of) $r(X^n)$ gives a bound on the Type-I error that is valid no matter what this stopping rule is. Standard (Neyman-Pearson) null hypothesis testing and tests derived from AIC-type methods are not robust in this sense. For example, such tests cannot be used if the stop-

ping rule is simply unknown, as is often the case when analyzing externally provided data — but this is just the tip of an iceberg of problems with nonrobust tests. For an exhaustive review of such problems we refer to Wagenmakers (2007) who builds on, amongst others, Berger and Wolpert (1988) and Pratt (1962).

Now, as first noted by Edwards et al. (1963), in simple versus composite testing (i.e. when $\mathcal{M}_0$ is a singleton), the output of BFMS, the Bayes factor, does provide a robust null hypothesis test. This is one of the main reasons why for example, in psychology, Bayesian testing is becoming more and more popular (Andrews and Baguley, 2012; Dienes, 2011), even among 'frequentist' researchers (Sanborn and Hills, 2014). Our third result, in Section 5.5, shows that the same holds for the switch criterion: if $\mathcal{M}_0$ is a singleton, so that the problem (5.1) reduces to a simple versus composite hypothesis test, then the evidence $r(X^n)$ associated with the switching criterion has the desired robustness property as well and thus in this sense behaves like the Bayes factor method. The advantage, from a frequentist point of view, of switching as compared to Bayes is then that switching is a lot more sensitive: our risk rate results directly imply that the Type II error (1 − power) of the switch criterion goes to 0 as soon as, at sample size n, the distance between the 'true' distribution μ_1 and the null model, i.e. $\inf_{\mu \in M_0} \|\mu - \mu_1\|_2^2$ is of order $(\log\log n)/n$; for Bayes factor testing, in order for the Type-II error to reach 0, this distance must be of order $(\log n)/n$ (this was informally recognized by Lhéritier and Cazals (2015), who reported substantially larger power of switching as compared to the Bayes factor method in a sequential two-sample testing setting).

Thus, switching gives us 'almost the best of three worlds': minimax rate optimality up to a $\log\log n$ factor (in contrast to BFMS), consistency (in contrast to AIC-type methods) and nonasymptotic insensitivity to optional stopping (in contrast to standard Neyman-Pearson testing) in combination with a small Type-II error.

Organization This paper is organized as follows. The switch criterion is introduced in Section 5.2. In Section 5.3, we provide some preliminaries: we list the loss/risk functions for which our result holds, describe the sets in which the truth is assumed to lie, and discuss the tension between consistency and rate-optimality. Suitable post-model-selection estimators to be used in combination with the switch criterion are introduced in Section 5.4, after which our main result on the worst-case risk of the switch criterion is stated. We also go into the relationship between the switch criterion and the Hannan-Quinn criterion in that section. In Section 5.5 we define robust null hypothesis tests, give some examples, and show that testing by switching has the desired nonasymptotic robustness to optional stopping; in constrast, AIC does not satisfy such a property at all and the Hannan-Quinn criterion only satisfies an asymptotic analogue. We also provide some simulations that illustrate our results. Section 5.6 provides some additional discussion and ideas for future work. All proofs are given in Section 5.7.

Notations and conventions We use $x^n = x_1, \ldots, x_n$ to denote n observations, each taking values in a sample space $\mathcal{X}$. For a set of parameters M, $\mu \in M$, and $x \in \mathcal{X}$, $p_\mu(x)$ invariably denotes the density or mass function of x under the distribution $\mathbb{P}_\mu$ of random variable X, taking values in $\mathcal{X}$. This is extended to n outcomes by independence, so that $p_\mu(x^n) := \prod_{i=1}^n p_\mu(x_i)$ and $\mathbb{P}_\mu(X^n \in A_n)$, abbreviated to $\mathbb{P}_\mu(A_n)$, denotes the probability

that $X^n \in A_n$ for $X^n = X_1, \ldots, X_n$ i.i.d. $\sim \mathbb{P}_\mu$. Similarly, $\mathbb{E}_\mu$ denotes expectation under $\mathbb{P}_\mu$. As is customary, we write $a_n \asymp b_n$ to denote $0 < \lim_{n\to\infty} \inf a_n/b_n \leq \lim_{n\to\infty} \sup a_n/b_n < \infty$. For notational simplicity we assume throughtout this paper that whenever we refer to a sample size n, then $n \geq 3$ to ensure that $\log\log n$ is defined and positive.

Throughout the text, we refer to standard properties of exponential families without always giving an explicit reference; all desired properties can be found, in precise form, in (Barndorff-Nielsen, 1978) and, on a less formal level, in (Grünwald, 2007, Chapter 18,19).

5.2 Model selection by switching

The *switch distribution* (Van Erven et al., 2012; 2007) is a modification of the Bayesian predictive distribution, inspired by Dawid's (1984) 'prequential' approach to statistics and the related *Minimum Description Length* (MDL) Principle (Barron et al., 1998; Grünwald, 2007). The corresponding *switch criterion* can be thought of as Bayes factor model selection with a prior on meta-models, where each meta-model consists of a sequence of basic models and associated starting times: until time t_1, follow model k_1, from time t_1 to t_2, follow model k_2, and so on. The fact that we only need to select between two nested parametric models allows us to considerably simplify the set-up of Van Erven et al. (2012), who dealt with countably infinite sets of arbitrary models.

It is convenient to directly introduce the switch criterion as a modification of the Bayes factor model selection (BFMS). Assuming equal prior $1/2$ on each of the models $\mathcal{M}_0$ and $\mathcal{M}_1$, BFMS associates each model $\mathcal{M}_k$, $k \in \{0,1\}$, with a *marginal distribution* $p_{B,k}$ with

$$p_{B,k}(x^n) := \int_{\mu\in M_k} \omega_k(\mu) p_\mu(x^n) d\mu, \tag{5.2}$$

where ω_k is a prior density on M_k. It then selects model $\mathcal{M}_1$ if and only if $p_{B,1}(x^n) > p_{B,0}(x^n)$.

The basic idea behind MDL model selection is to generalize this in the sense that each model $\mathcal{M}_k$ is associated with *some* 'universal' distribution $p_{U,k}$; one then picks the k for which $p_{U,k}(x^n)$ is largest. $p_{U,k}$ may be set to the Bayesian marginal distribution, but other choices may be preferable in some situations. Switching is an instance of this; in our simplified setting, it amounts to associating $\mathcal{M}_0$ with a Bayes marginal distributon $p_{B,0}$ as before. $p_{U,1}$ however is set to the *switch distribution* $p_{\mathrm{sw},1}$. This distribution corresponds to a switch between models $\mathcal{M}_0$ and $\mathcal{M}_1$ at some sample point s, which is itself uncertain; before point s, the data are modelled as coming from $\mathcal{M}_0$, using $p_{B,0}$; after point s, they are modelled as coming from $\mathcal{M}_1$, using $p_{B,1}$. Formally, we denote the strategy that switches from the simple to the complex model after t observations by $\bar{p}_t$; $p_{\mathrm{sw},1}$ is then defined as the marginal distribution by averaging $\bar{p}_t$ over t, with some probability mass function π (analogous to a Bayesian prior) over $t \in \{1,2,\ldots\}$:

$$\bar{p}_t(x^n) = p_{B,0}(x^{t-1}) \cdot p_{B,1}(x_t, \ldots, x_n \mid x^{t-1})$$
$$p_{\mathrm{sw},1}(x^n) = \sum_{t=1}^{\infty} \pi(t) \bar{p}_t(x^n),$$

where switching at $t = 1$ corresponds to predicting with $p_{B,1}$ at each data point, and switching at any $t > n$ to predicting with $p_{B,0}$. We remind the reader that even for i.i.d. models, $p_{B,1}(x_t, \ldots, x_n \mid x^{t-1})$ usually depends on x^{t-1} — the Bayes predictive distribution learns from data. The model selection criterion δ_{sw} mapping sequences of arbitrary length to $k \in \{0, 1\}$ is then defined, for each n, as follows:

$$\delta_{\mathrm{sw}}(x^n) = \begin{cases} 0 & \text{if } \dfrac{p_{\mathrm{sw},1}(x^n)}{p_{\mathrm{B},0}(x^n)} \leq 1 \\ 1 & \text{if } \dfrac{p_{\mathrm{sw},1}(x^n)}{p_{\mathrm{B},0}(x^n)} > 1 \end{cases} . \tag{5.3}$$

When defining $p_{\mathrm{sw},1}$ it is sufficient to consider switching times that are equal to a power of two. Thus, we restrict attention to 'priors' π on switching time with support on $2^0, 2^1, 2^2, \ldots$. For our subsequent results to hold, π should be such that $\pi(2^i)$ decays like $i^{-\kappa}$ for some $\kappa \geq 2$. An example of such a prior with $\kappa = 2$ is $\pi(2^i) = 1/((i+1)(i+2))$, $\pi(j) = 0$ for any j that is not a power of 2.

To prepare for Theorem 5.5, we instantiate the switch criterion to the problem (5.1). We define $p_{B,1}$ as any distribution of the form (5.2) where ω_1 is a continuous prior density on M_1 that is strictly positive on all $\mu \in M_1$. To define $p_{B,0}$ we need to take a slight detour, because we parameterized $\mathcal{M}_0$ in terms of an M_0 that has a fixed value on its final $m_1 - m_0$ components: it is an m_0-dimensional family with an m_1-dimensional parameterization, so one cannot easily express a prior on $\mathcal{M}_0$ as a density on M_0. To overcome this, we distinguish between the case that $m_0 = 0$ and $m_0 > 0$. In the former case M_0 has a single element ν, and we define $p_{B,0} = p_\nu$. In the latter case, we define $\Pi'_0 : M_0 \to \mathbb{R}^{m_0}$ as the projection of $\mu \in M_0$ on its first m_0 components, and $\Pi'_0(M_0) := \{\Pi'_0(\mu) : \mu \in M_0\}$. For $\mu \in M_0$, we define $p_{\Pi'_0(\mu)} = p_\mu$, and we then let ω_0 be a continuous strictly positive prior density on $\Pi'_0(M_0)$, and we define $p_{B,0}(x^n) := \int_{\mu' \in \Pi'_0(M_0)} \omega_0(\mu') p_{\mu'}(x^n) d\mu'$.

Two important remarks are in order: first, the fact that we associate $\mathcal{M}_1$ with a distribution incorporating a 'switch' from $\mathcal{M}_0$ to $\mathcal{M}_1$ *does not mean* that we really believe that data were sampled, until some point t, according to $\mathcal{M}_0$ and afterwards according to $\mathcal{M}_1$. Rather, it is suggested by prequential and MDL considerations, which suggest that one should pick the model that performs best in sequentially predicting data; and if the data are sampled from a distribution in $\mathcal{M}_1$ that is not in $\mathcal{M}_0$, but quite close to it in KL divergence, then $p_{\mathrm{B},1}$ is suboptimal for sequential prediction, and can be substantially outperformed by $p_{\mathrm{sw},1}$. This is explained at length by Van Erven et al. (2012), and Figure 1 in that paper especially illustrates the point. The same paper also explains how one can use dynamic programming to arrive at an implementation that has the same computational efficiency as computation of the standard Bayes model selection decision.

Second, the criterion (5.3) as defined here is not 100% equivalent to the special case of the construction of Van Erven et al. (2012) specialized to two models, but rather an easily-explained simplification thereof. Yet, all our results continue to hold if we were to follow the original construction, as explained in Section 5.7.5; and conversely, the strong consistency result for the construction of Van Erven et al. (2012) trivially continues to hold for the criterion (5.3) used in the present paper.

5.3 Rate-optimality of post-model selection estimators

This section contains some background to our main result, Theorem 5.5. In Section 5.3.1, we first list the loss functions for which our main result holds, and define the CINECSI sets in which the truth assumed to lie. We then discuss the minimax parametric risk for our model selection problem in Section 5.3.2. This section ends with a discussion on the generality of the impossibility result of Yang (2005) in Section 5.3.3.

5.3.1 Loss functions and CINECSI sets

Let $\mathcal{M} = \{p_\mu \mid \mu \in M\}$ be an exponential family given in its mean-value parameterization with $M \subset \mathbb{R}^m$ a product of m open, possibly but not necessarily unbounded intervals for some $m > 0$; see Section 5.7 for a formal definition of exponential families and mean-value parameterizations. Note that we do not require the family to be 'full'; for example, the Bernoulli model with success probability $\mu \in M_1 = (0.2, 0.4)$ counts as an exponential family in our (standard) definition.

Suppose that we measure the quality of a density $p_{\mu'}$ as an approximation to p_μ by a loss function $L : M \times M \to \mathbb{R}$. The standard definition of the (instantaneous) *risk* of estimator $\breve{\mu} : \bigcup_{i>0} \mathcal{X}^i \to M$ at sample size n, as defined relative to loss L, is given by its expected loss,

$$R(\mu, \breve{\mu}, n) = \mathbb{E}_\mu \left[L(\mu, \breve{\mu}(X^n)) \right],$$

where $\mathbb{E}_\mu$ denotes expectation over $X_1, \ldots, X_n$ i.i.d. $\sim \mathbb{P}_\mu$. Popular loss functions are:

1. The *squared error loss*: $d_{SQ}(\mu', \mu) = \|\mu' - \mu\|_2^2$;
2. The *standardized squared error loss* which is a version of the *squared Mahalonobis distance*, defined as

$$d_{ST}(\mu' \| \mu) := (\mu - \mu')^T I(\mu')(\mu - \mu'), \tag{5.4}$$

   where T denotes transpose, $I(\cdot)$ is the Fisher information matrix, and we view μ and μ' as column vectors;
3. The Rényi divergence of order $1/2$, defined as

$$d_R(\mu', \mu) = -2 \log \mathbb{E}_{\mu'} \left[\left(p_\mu(X)/p_{\mu'}(X) \right)^{\frac{1}{2}} \right];$$

4. The squared Hellinger distance $d_{H^2}(\mu', \mu) = 2 \left(1 - \mathbb{E}_{\mu'} \left[\left(p_\mu(X)/p_{\mu'}(X) \right)^{\frac{1}{2}} \right] \right)$;
5. The KL (Kullback-Leibler) divergence $D(p_{\mu'} \| p_\mu)$, henceforth abbreviated to $D(\mu' \| \mu)$.

We note that there is a direct relationship between the Rényi divergence and squared Hellinger distance:

$$d_{H^2}(\mu', \mu) = 2 \left(1 - e^{-d_R(\mu', \mu)/2} \right). \tag{5.5}$$

In fact, as we show below, these loss functions are all equivalent (equal up to universal constants) on CINECSI sets. Such sets will play an important role in the sequel. They are defined as follows:

Definition 5.1 (CINECSI)**.** A *CINECSI (Connected, Interior-Non-Empty-Compact-Subset-of-Interior) subset* of a set M is a connected subset of the interior of M that is itself compact and has nonempty interior.

The following proposition is proved in Section 5.7.

Proposition 5.2. Let M be the mean-value parameter space of an exponential family as above, and let M' be an CINECSI subset of M. Then there exist positive constants $c_1, c_2, \ldots, c_6$ such that for all $\mu, \mu' \in M'$,

$$c_1\|\mu' - \mu\|_2^2 \leq c_2 \cdot d_{ST}(\mu'\|\mu) \leq d_{H^2}(\mu', \mu) \leq d_R(\mu', \mu) \leq D(\mu'\|\mu) \leq c_3\|\mu' - \mu\|_2^2. \quad (5.6)$$

and for all $\mu' \in M', \mu \in M$ (i.e. μ is now not restricted to lie in M'),

$$d_{H^2}(\mu', \mu) \leq c_4\|\mu' - \mu\|_2^2 \leq c_5 \cdot d_{ST}(\mu'\|\mu) \leq c_6\|\mu' - \mu\|_2^2. \quad (5.7)$$

CINECSI subsets are a variation on the INECCSI sets of (Grünwald, 2007). Our main result, Theorem 5.5, holds for all of the above loss functions, and for general 'sufficiently efficient' estimators. While the equivalence of the losses above on CINECSI sets is a great help in the proofs, we emphasize that we never require these estimators to be restricted to CINECSI subsets of M — although, since we require M to be open, every 'true' $\mu \in M$ will lie in *some* CINECSI subset M' of M, a statistician who employs the model $\mathcal{M}$ cannot know what this M' is, so such a requirement would be unreasonably strong.

5.3.2 Minimax parametric risk

We say that a quantity f_n converges at rate g_n if $f_n \asymp g_n$. We say that an estimator $\breve{\mu}$ is *minimax rate optimal* relative to a model $\mathcal{M} = \{p_\mu \mid \mu \in M\}$ *restricted to a subset* $M' \subset M$ if

$$\sup_{\mu \in M'} R(\mu, \breve{\mu}, n)$$

converges at the same rate as

$$\inf_{\dot{\mu}} \sup_{\mu \in M'} R(\mu, \dot{\mu}, n), \quad (5.8)$$

where $\dot{\mu}$ ranges over all estimators of μ at sample size n, that is, all measurable functions from $\mathcal{X}^n$ to M.

For most parametric models encountered in practice, the minimax risk (5.8) is of order $1/n$ when R is defined relative to any of the loss measures defined in Section 5.3.1 and M' is an arbitrary CINECSI subset of M (Van der Vaart, 1998). In particular this holds if $\mathcal{M}$ is an exponential family. For this reason, from now on we refer to $1/n$ as the *minimax parametric rate*. Note that, crucially, the restriction $\mu \in M'$ is imposed only on the data-generating distribution, not on the estimators, and, since we will require models with open parameter sets M such that for every $\delta > 0$, there is a CINECSI subset M'_δ of M with $\sup_{\mu \in M} \inf_{\mu' \in M'_\delta} \|\mu - \mu'\|_2^2 < \delta$, every possible $\mu \in M$ will also lie in some CINECSI subset M'_δ that 'nearly' covers M_δ. This makes the restriction to CINECSI M' in the definition above a mild one. Still, it is necessary: at least for the squared error loss, for most exponential families (the exception being the Gaussian location family), we have

$\inf_{\dot{\mu}} \sup_{\mu \in M'_\delta} R(\mu, \dot{\mu}, n) = C_\delta / n$ for some constant $C_\delta > 0$, but the smallest constant for which this holds may grow arbitrarily large as $\delta \to 0$, the reason being that the determinant of the Fisher information may tend to 0 or ∞ as $\delta \to 0$.

Now consider a model selection criterion $\delta : \bigcup_{i>0} \mathcal{X}^i \to \{0, 1, \ldots, K-1\}$ that selects, for given data x^n of arbitrary length n, one of a finite number K of parametric models $\mathcal{M}_0, \ldots, \mathcal{M}_{K-1}$ with respective parameter sets $M_0, \ldots, M_{K-1}$. One way to evaluate the quality of δ is to consider the risk attained after first selecting a model and then estimating the parameter vector μ using an estimator $\breve{\mu}_k$ associated with each model $\mathcal{M}_k$. This *post-model selection estimator* (Leeb and Pötscher, 2005) will be denoted by $\breve{\mu}_{\breve{k}}(x^n)$, where $\breve{k}$ is the index of the model selected by δ. The risk of a model selection criterion δ is thus $R(\mu, \delta, n) = \mathbb{E}_\mu \left[L(\mu, \breve{\mu}_{\breve{k}}(X^n)) \right]$, where L is a given loss function, and its worst-case risk relative to μ restricted to $M'_k \subset M_k$ is given by

$$\sup_{\mu \in M'_k} R(\mu, \delta, n) = \sup_{\mu \in M'_k} \mathbb{E}_\mu \left[L(\mu, \breve{\mu}_{\breve{k}}(X^n)) \right]. \tag{5.9}$$

We are now ready to define what it means for a model selection criterion to achieve the minimax parametric rate.

Definition 5.3. A model selection criterion δ *achieves the minimax parametric rate* if there exist estimators $\breve{\mu}_k$, one for each $\mathcal{M}_k$ under consideration, such that, for *every* CINECSI subset M'_k of M:

$$\sup_{\mu \in M'_k} R(\mu, \delta, n) \asymp 1/n.$$

Just as in the fixed-model case, the restriction $\mu \in M'_k$ is imposed only on the data-generating distribution, not on the estimators.

5.3.3 The result of Yang (2005) transplanted to our setting

In this paper, as stated in the introduction, we further specialize the setting above to problem (5.1) where we select between two nested exponential families, which we shall always assume to be given in their mean-value parameterization. To be precise, the 'complex' model $\mathcal{M}_1$ contains distributions from an exponential family parametrized by an m_1-dimensional mean vector μ, and the 'simple' model $\mathcal{M}_0$ contains distributions with the same parametrization, where the final $m_1 - m_0$ components are fixed to values $v_{m_0+1}, \ldots, v_{m_1}$. We introduce some notation to deal with the assumption that M_1 and its restriction of M_0 to its first m_0 components are products of open intervals. Formally, we require that M_1 and M_0 are of the form

$$\begin{aligned} M_1 &= (\zeta_{1,1}, \eta_{1,1}) \times \ldots \times (\zeta_{1,m_1}, \eta_{1,m_1}) \\ M_0 &= (\zeta_{0,1}, \eta_{0,1}) \times \ldots \times (\zeta_{0,m_0}, \eta_{0,m_0}) \times \{v_{m_0+1}\} \times \ldots \times \{v_{m_1}\} \end{aligned} \tag{5.10}$$

where, for $j = 1, \ldots, m_0$, we have $-\infty \le \zeta_{1,j} \le \zeta_{0,j} < \eta_{0,j} \le \eta_{1,j} \le \infty$; and for $j = m_0 + 1, \ldots, m_1$, we have $-\infty \le \zeta_{1,j} < v_j < \eta_{1,j} \le \infty$.

For example, $\mathcal{M}_1$ could contain all normal distributions with mean μ and variance σ^2, with mean value parameters $\mu_1 = \mu^2 + \sigma^2$ and $\mu_2 = \mu$, and $M_1 = (0, \infty) \times (-\infty, \infty)$, while

$\mathcal{M}_0$ could contain all normal distributions with mean zero and unknown variance σ^2, so $M_0 = (0, \infty) \times \{0\}$.

Yang (2005) showed in a linear regression context that a model selection criterion cannot both achieve the minimax optimal parametric rate and be consistent; a practitioner is thus forced to choose between a rate-optimal method such as AIC and a consistent method such as BIC. Inequality (5.12) below provides some insight into why this *AIC-BIC dilemma* can occur. A similar inequality appears in Yang's paper for his linear regression context, but it is still valid in our exponential family setting, and the derivation — which we now give — is essentially equivalent.

To state the inequality, we need to relate $\mu_1 \in M_1$ to a component in M_0. For any given $\mu_1 = (\mu_{1,1}, \ldots, \mu_{1,m_1})^T \in M_1$, we will define

$$\Pi_0(\mu_1) := (\mu_{1,1}, \ldots, \mu_{1,m_0}, \nu_{m_0+1}, \ldots, \nu_{m_1})^T \tag{5.11}$$

to be the *projection* of μ_1 on M_0. The difference between Π_0 of (5.11) and Π_0' in Section 5.2 is that Π_0 is a function from $\mathbb{R}^{m_1}$ to $\mathbb{R}^{m_1}$, whereas Π_0' is a function from $\mathbb{R}^{m_1}$ to $\mathbb{R}^{m_0}$; $\Pi_0(\mu_1)$ and $\Pi_0'(\mu_1)$ agree in the first m_0 components. Note that $\Pi_0(\mu_1)$ obviously minimizes, among all $\mu \in M_0$, the squared Euclidean distance $\|\mu - \mu_1\|_2^2$ to p_{μ_1}; somewhat less obviously it also minimizes, among $\mu \in M_0$, the KL divergence $D(p_{\mu_1} \| p_\mu)$ (Grünwald, 2007, Chapter 19); we may thus think of it as the 'best' approximation of the 'true' μ_1 within M_0; we will usually abbreviate $\Pi_0(\mu_1)$ to μ_0.

Let A_n be the event that the complex model is selected at sample size n. Since $\mathcal{M}_1$ is an exponential family, the MLE $\widehat{\mu}_1$ is unbiased and $\widehat{\mu}_0$ coincides with $\widehat{\mu}_1$ in the first m_0 components, so that $\mathbb{E}_{\mu_1}\left[\mu_0 - \widehat{\mu}_0(X^n)\right] = 0$, and hence we can rewrite, for any $\mu_1 \in M_1$, the squared error risk as

$$\begin{aligned}
R(\mu_1, \delta, n) &= \mathbb{E}_{\mu_1}\left[\mathbf{1}_{A_n}\|\mu_1 - \widehat{\mu}_1(X^n)\|_2^2 + \mathbf{1}_{A_n^c}\|\mu_1 - \widehat{\mu}_0(X^n)\|_2^2\right] \\
&= \mathbb{E}_{\mu_1}\left[\mathbf{1}_{A_n}\|\mu_1 - \widehat{\mu}_1(X^n)\|_2^2 + \mathbf{1}_{A_n^c}\|\mu_0 - \widehat{\mu}_0(X^n)\|_2^2 + \mathbf{1}_{A_n^c}\|\mu_1 - \mu_0\|_2^2\right] \\
&\leq \mathbb{E}_{\mu_1}\left[\|\mu_1 - \widehat{\mu}_1(X^n)\|_2^2 + \|\mu_0 - \widehat{\mu}_0(X^n)\|_2^2\right] + \mathbb{P}(A_n^c)\|\mu_1 - \mu_0\|_2^2 \\
&\leq 2R(\mu_1, \widehat{\mu}_1, n) + \mathbb{P}(A_n^c)\|\mu_1 - \mu_0\|_2^2.
\end{aligned} \tag{5.12}$$

The first part of the proof of our main result, Theorem 5.5, extends this decomposition to general estimators and loss functions.

The first term on the right of (5.12) is of order $1/n$. The second term depends on the 'Type-II error', i.e. the probability of selecting the simple model when it is not actually true. A low worst-case risk is attained if this probability is small, even if the true parameter is close to μ_0. This does leave the possibility for a risk-optimal model selection criterion to incorrectly select the complex model with high probability. In other words, a risk-optimal model selection method may not be consistent if the simple model is correct. The theorem by Yang (2005), arguing from decomposition (5.12), essentially demonstrates that it cannot be. Due to the general nature of (5.12), it seems likely that his result holds in much more general settings: a procedure attains a low worst-case risk by selecting the complex model with high probability, which is excellent if the complex model is indeed true, but leads to inconsistency if the simple model is correct. Indeed, we have shown in earlier work that the dilemma is not restricted to linear regression, but occurs in our exponential family

problem (5.1) as well as long as $\mathcal{M}_0 = \{\nu\}$ is a singleton (see Van der Pas (2013) for the proof, which is a simple adaptation of Yang's proof that, we suspect, can be extended to nonsingleton $\mathcal{M}_0$ as well). Hence, as the switch criterion is strongly consistent Van Erven et al. (2012), we know that the worst-case risk rate of the switch criterion cannot be of the order $1/n$ in general.

5.4 Main result

We perform model selection by using the switch criterion, as specified in Section 5.2. After the model selection, we estimate the underlying parameter μ. We discuss post-model selection estimators suitable to our problem in Section 5.4.1. We are then ready to present our main result, Theorem 5.5 in Section 5.4.2, stating that the worst-case risk for the switch criterion under the loss functions listed in Section 5.3.1 attains the minimax parametric rate up to a $\log\log n$ factor.

5.4.1 Post-model selection: sufficiently efficient estimators

Our goal is to determine the worst-case rate for the switch criterion applied to two nested exponential families, which we combine with an estimator as follows: if the simple model is selected, μ will be estimated by an estimator $\check{\mu}_0$ with range M_0. If the complex model is selected, the estimate of μ will be provided by another estimator $\check{\mu}_1$ with range M_1. Our result will hold for all estimators $\check{\mu}_0$ and $\check{\mu}_1$ that are *sufficiently efficient*:

Definition 5.4 (sufficiently efficient)**.** The estimators $\{\check{\mu}_k \to M_k \mid k \in \{0,1\}\}$ are *sufficiently efficient* with respect to a divergence measure $d_{\mathrm{gen}}(\cdot\|\cdot)$ if (with $\mu_0 = \Pi_0(\mu_1)$ as in (5.11)), for every CINECSI subset M_1' of M_1, there exists a constant $C > 0$ such that for all n,

$$\sup_{\mu_1 \in M_1'} \mathbb{E}_{\mu_1}[d_{\mathrm{gen}}(\mu_0\|\check{\mu}_0)] \le C \cdot \sup_{\mu_1 \in M_1'} \mathbb{E}_{\mu_1}[d_{\mathrm{gen}}(\mu_1\|\check{\mu}_1)] \le \frac{C}{n}. \tag{5.13}$$

Example 1. [Sufficient efficiency for MLE's for squared (standardized) error and Hellinger] If $\mathcal{M}_0$ and $\mathcal{M}_1$ are exponential families given in their mean-value parameterization with $M_0 \subset M_1$ as in (5.10), then $\mathcal{M}_1$ has sufficient statistic $\varphi \equiv (\varphi_1, \ldots, \varphi_{m_1})^T : \mathcal{X} \to \mathbb{R}^{m_1}$ (see Section 5.7 for the formal definition). Now by standard properties of exponential families, if

$$n^{-1} \sum \varphi(X_i) \in M_1, \tag{5.14}$$

then the ML estimator for model $\mathcal{M}_k$ is equal to $n^{-1}\sum_{i=1}^n \varphi(X_i)$. For many full families such as the full (multivariate) Gaussians, Gamma and many others, (5.14) holds μ-almost surely for each n, for all $\mu \in M_1$. Then the MLE is almost surely well-defined for M_1. We can then take $\check{\mu}_1 := \widehat{\mu}_1$ to be the MLE for $\mathcal{M}_1$, and $\check{\mu}_0$ to be its projection on the first m_0 coordinates (usually (5.14) will still hold for M_0 and then this $\check{\mu}_0$ will be the MLE for $\mathcal{M}_0$). This pair of estimators will be sufficiently efficient for (standardized) squared error and squared Hellinger distance, i.e. (5.13) holds for these three losses. To show this, note that from (5.27) in Proposition 5.2, we see that it is sufficient to show that (5.13) holds for the squared error loss. Since the j-th component of $\widehat{\mu}_1$ is equal to $n^{-1}\sum_{i=1}^n \varphi_j(X_i)$ and

$\mathbb{E}_{\mu_1}[n^{-1}\sum_{i=1}^n \varphi_j(X_i)] = \mu_{1,j}$ and $\text{VAR}_{\mu_1}\left[n^{-1}\sum_{i=1}^n \varphi_j(X_i)\right] = n^{-1}\text{VAR}_{\mu_1}\left[\varphi_j(X_1)\right]$, it suffices to show that

$$\sup_{\mu_1 \in M_1'} \sup_{j=1,\ldots,m_1} \text{VAR}_{\mu_1}\left[\varphi_j(X_1)\right] = O(1),$$

which is indeed the case since M_1' is an CINECSI set, so that the variance of all φ_j's is uniformly bounded on M_1' (Barndorff-Nielsen, 1978).

Example 2. [Other sufficiently efficient estimators for squared (standardized) error and Hellinger] For models such as the Bernoulli or multinomial, (5.14) may fail to hold with positive probability: the full Bernoulli exponential family does not contain the distributions with $P(X_1 = 1) = 1$ and $P(X_1 = 0) = 1$, so if after n examples, only zeros or only ones have been observed, the MLE is undefined. We can then go either of three ways. The first way, which we shall not pursue in detail here, is to work with so-called 'aggregate' exponential families, which are extensions of full families to their limit points. For models with finite support (such as the multinomial) these are well-defined (Barndorff-Nielsen, 1978, page 154–158) and then the MLE's for these extended families are almost surely well-defined again, and the MLE's are sufficiently efficient by the same reasoning as above. Another approach that works in some cases (e.g. multinomial) is to take $\breve{\mu}_1$ to be a truncated MLE, that, at sample size n, maps X^n to the MLE within some CINECSI subset $M_1^{(n)}$ of M_1, where $M_1^{(n)}$ converges to M_1 as n increases in the sense that $\sup_{\mu \in M_1^{(n)}, \mu' \in M_1 \setminus M_1^{(n)}} \|\mu - \mu'\|_2^2 = O(1/n)$. The resulting truncated MLE, and its projection on M_0 (usually itself a truncated MLE) will then again be sufficiently efficient. This approach also works if the models $\mathcal{M}_0$ and $\mathcal{M}_1$ are not full but restricted families to begin with. For full families though, a more elegant approach than truncating MLE's is to work with Bayesian posterior MAP estimates with conjugate priors. For steep exponential families (nearly all families one encounters in practice are steep), one can always find conjugate priors such that the Bayes MAP estimates based on these priors exist and take a value in M_1 almost surely (Grünwald and de Rooij, 2005). They then take the form $\breve{\mu}_1 = \sum_{i=1}^n (\varphi(X_i) + \lambda_0 \mu_1^\circ)/(n + \lambda_0)$, where $\lambda_0 > 0$ and $\mu_1^\circ \in M_1$ are determined by the prior. $\breve{\mu}_0$ can then again be taken to be the projection of $\breve{\mu}_1$ onto M_0. Under the assumption that μ_1 is contained in a CINECSI set M_1', one can now again show, using the same arguments as in Example 1, that such estimators are sufficiently efficient for squared (standardized) error and Hellinger loss.

Example 3. [Sufficient efficiency for Rényi and KL divergence] As is well-known, for the multivariate Gaussian model with fixed covariance matrix, the squared error risk and KL divergence are identical up to constant factors, so the unrestricted MLE's will still be sufficiently efficient for KL divergence. For other models, though, the MLE will not always be sufficiently efficient. For example, with the Bernoulli model and other models with finite support, to make the unrestricted MLE's well-defined, we would have to extend the family to its boundary points as indicated in Example 1. Since, however, for any $0 < \mu < 1$ and $\mu' = 0$, the KL divergence $D(\mu\|\mu') = \infty$ and $\mathbb{P}_\mu(\hat{\mu}(X^n) = \mu') > 0$, the unrestricted MLE in the full Bernoulli model including the boundaries will have infinite risk and thus will not be sufficiently efficient. The MAP estimators tend to behave better though: Grünwald and de Rooij (2005) implicitly show that for 1-dimensional families, under weak conditions on the family (Condition 1 underneath Theorem 1 in their paper) —

which were shown to hold for a number of families such as Bernoulli, Poisson, geometric — sufficient efficiency for the KL divergence still holds for MAP estimators of the form above. We conjecture that a similar result can be shown for multidimensional families, but will not attempt to do so here.

5.4.2 Main result: risk of the switch criterion

We now present our main result, which states that for the exponential family problem under consideration, the worst-case instantaneous risk rate of δ_{sw} is of order $(\log\log n)/n$. Hence, the worst-case instantaneous risk of δ_{sw} is very close to the lower bound of $1/n$, while the criterion still maintains consistency.

The theorem holds for any of the loss functions listed in Section 5.3.1. We denote this by using the generic loss function d_{gen}, which can be one of the following loss functions: squared error loss, standardized squared error los, KL divergence, Rényi divergence of order 1/2, or squared Hellinger distance.

Theorem 5.5. Let $\mathcal{M}_0 = \{p_\mu \mid \mu \in M_0\}$ and $\mathcal{M}_1 = \{p_\mu \mid \mu \in M_1\}$ be nested exponential families in their mean-value parameterization, where $M_0 \subseteq M_1$ are of the form (5.10). Assume:

1. $\breve{\mu}_0$ and $\breve{\mu}_1$ are sufficiently efficient estimators relative to the chosen loss d_{gen};
2. δ_{sw} is constructed with $p_{B,0}$ and $p_{B,1}$ defined as in Section 5.2 with priors ω_k that admit a strictly positive, continuous density;
3. and $p_{\mathrm{sw},1}$ is defined relative to a prior π with support on $\{0,1,2,4,8,\ldots\}$ and $\pi(2^i) \propto i^{-\kappa}$ for some $\kappa \geq 2$.

Then for every CINECSI subset M_1' of M_1, we have:

$$\sup_{\mu_1 \in M_1'} R(\mu_1, \delta_{\mathrm{sw}}, n) = O\left(\frac{\log\log n}{n}\right),$$

for $R(\mu, \delta_{\mathrm{sw}}, n)$ the risk at sample size n defined relative to the chosen loss d_{gen}.

Example 4. [Our setting vs. Yang's] Yang (2005) considers model selection between two nested linear regression models with fixed design, where the errors are Gaussian with fixed variance. The risk is measured as the in-model squared error risk ('in-model' means that the loss is measured conditional on a randomly chosen design point that already appeared in the training sample). Within this context he shows that every model selection criterion that is (weakly) consistent cannot achieve the $1/n$ minimax rate. The exponential family result above leads one to conjecture that the switch distribution achieves $O((\log\log n)/n)$ risk in Yang's setting as well. We suspect that this is so, but actually showing this would require substantial additional work. Compared to our setting, Yang's setting is easier in some and harder in other respects: under the fixed-variance, fixed design regression model, the Fisher information is constant, making asymptotic results hold nonasymptotically, which would greatly facilitate our proofs (and obliterate any need to consider CINECSI sets or

undefined MLE's). On the other hand, evaluating the risk conditional on a design point is not something that can be directly embedded in our proofs.

Example 5. [Switching vs. Hannan-Quinn] In their comments on Van Erven et al. (2012), Lauritzen (2012) and Cavanaugh (2012) suggested a relationship between the switch model selection criterion and the criterion due to Hannan and Quinn (1979). For the exponential family models under consideration, the Hannan-Quinn criterion with parameter c, denoted as HQ, selects the simple model, i.e. $\delta_{\mathrm{HQ}}(x^n) = 0$, if

$$-\log p_{\widehat{\mu}_0}(x^n) < -\log p_{\widehat{\mu}_1}(x^n) + c \log\log n,$$

and the complex model otherwise. In their paper, Hannan and Quinn show that this criterion is strongly consistent for $c > 1$.

As shown by Barron et al. (1999), under some regularity conditions, penalized maximum likelihood criteria achieve worst-case quadratic risk of the order of their penalty divided by n. One can show (details omitted) that this is also the case in our specific setting and hence, that the worst-case risk rate of HQ for our problem is of order $(\log\log n)/n$. Our main result, Theorem 5.5, shows that the same risk rate is achieved by the switch distribution, thus partially confirming the conjecture of Lauritzen (2012) and Cavanaugh (2012): HQ achieves the same risk rate as the switch distribution and, for the right choice of c, is also strongly consistent. This suggests that the switch distribution and HQ, at least for some specific value c_0, may behave asymptotically indistinguishably. The earlier results of Van der Pas (2013) suggest that this is indeed the case if $\mathcal{M}_0$ is a singleton; if $\mathcal{M}_0$ has dimensionality larger than 0, this appears to be a difficult question which we will not attempt to resolve here — in this sense the conjecture of Lauritzen (2012) and Cavanaugh (2012) has only been partially resolved.

Because HQ and δ_{sw} have been shown to be both strongly consistent and achieve the same rates for this problem, one may wonder whether one criterion is to be preferred over the other. For this parametric problem, HQ has the advantage of being simpler to analyze and implement. The criterion δ_{sw} can however, be used to define a robust hypothesis test as in Section 5.5 below. As we shall see there, HQ is insensitive to optional stopping in an asymptotic sense only, whereas robust tests such as the switch criterion are insensitive to optional stopping in a much stronger, nonasymptotic sense. Except for the normal location model, for which the asymptotics are precise, the HQ criterion cannot be easily adapted to define such a robust, nonasymptotic test. Another advantage of switching is that it can be combined with arbitrary priors and applied much more generally, for example when the constituting models are themselves nonparametric (Lhéritier and Cazals, 2015), are so irregular that standard asymptotics such as the law of the iterated logarithm are no longer valid, or are represented by black-box predictors such that ML estimators and the like cannot be calculated. In all of these cases the switch criterion can still be defined and — given the explanation in the introduction of Van Erven et al. (2012) — one may still expect it to perform well.

5.5 Robust null hypothesis tests

Bayes factor model selection, the switch criterion, AIC, BIC, HQ and most model selection methods used in practice are really based on thresholding the output of a more informative *model comparison method*. This is defined as a function from data of arbitrary size to the nonnegative reals. Given data x^n, it outputs a number $r(x^n)$ between 0 and ∞ that is a deterministic function of the data x^n. Every model comparison method r and threshold t has an associated model selection method $\delta_{r,t}$ that outputs 1 (corresponding to selecting model $\mathcal{M}_1$) if $r(x^n) \leq t$, and 0 otherwise. As explained below, such model comparison methods can often be viewed as performing a null hypothesis with $\mathcal{M}_0$ the null hypothesis, $\mathcal{M}_1$ the alternative hypothesis and t akin to a significance level.

Example 1 (BFMS): The output of the Bayes factor model comparison method is the posterior odds ratio $r_{\text{Bayes}}(x^n) = \mathbb{P}(\mathcal{M}_0|x^n)/\mathbb{P}(\mathcal{M}_1|x^n)$. The associated model selection method (BFMS) with threshold t selects model $\mathcal{M}_1$ if and only $r_{Bayes}(x^n) \leq t$.

Example 2 (AIC): Standard AIC selects model $\mathcal{M}_1$ if $\log(p_{\widehat{\mu}_1}(x^n)/p_{\widehat{\mu}_0}(x^n)) > m_1 - m_0$. We may however consider more conservative versions of AIC that only select $\mathcal{M}_1$ if

$$\log(p_{\widehat{\mu}_1}(x^n)/p_{\widehat{\mu}_0}(x^n)) - (m_1 - m_0) \geq -\log t. \tag{5.15}$$

We may thus think of AIC as a model comparison method that outputs the left-hand side of (5.15), and that becomes a model selection method when supplied with a particular t.

Now classical Neyman-Pearson null hypothesis testing requires the *sampling plan*, or equivalently, the *stopping rule*, to be determined in advance to ensure the validity of the subsequent inference. In the important special case of (generalized) likelihood ratio tests, this even means that the sample size n has to be fixed in advance. In practice, greater flexibility in choosing the sample size n is desirable (Wagenmakers (2007) provides sophisticated examples and discussion). Below, we discuss hypothesis tests that allow such flexibility by virtue of the property that their Type I-error probability remains bounded irrespective of the stopping rule used. These *robust* null hypothesis tests are defined below. As will be shown, whenever the null hypothesis $\mathcal{M}_0 = \{p_{\mu_0}\}$ is 'simple', i.e. a singleton (simple vs. composite testing), both Bayes factor model selection (BFMS) and the switch distribution define such robust null hypothesis tests, whereas AIC does not and HQ does so only in an asymptotic sense. As we argue in Section 5.5.3, the advantage of switching over BFMS is then that, while both share the robustness Type-I error property, switching has significantly smaller Type-II error (larger power) than BFMS when the 'truth' is close to $\mathcal{M}_0$, which is a direct consequence of it having a smaller risk under the alternative $\mathcal{M}_1$. To make this point concrete, and to indicate what may happen if $\mathcal{M}_0$ is not a singleton, we provide a simulation study in Section 5.5.4.

5.5.1 Bayes factors with singleton $\mathcal{M}_0$ are robust under optional stopping

In many cases, for each $0 < \alpha < 1$ there is an associated threshold $t(\alpha)$, which is a strictly increasing function of α, such that for every $t \leq t(\alpha)$ we have that $\delta_{r,t}$ becomes a null hypothesis significance test (NHST) with type-I error probability bounded by α. In particular, then $\delta_{r,t(\alpha)}$ is a standard NHST with type-I error bounded by α. For example,

for AIC with $M_0 = \{0\}$ and $M_1 = \mathbb{R}$ representing the normal family of distributions with unit variance, we may select $t(\alpha) = \exp(-2/z^2_{\alpha/2})$, where $z_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of the standard normal distribution. This results in the generalized likelihood ratio test at significance level α.

We say that model comparison method r defines a *robust null hypothesis test* for null hypothesis $\mathcal{M}_0$ and significance level α if for all $\mu_0 \in M_0$,

$$P_{\mu_0}(\exists n : \delta_{r,t(\alpha)}(X^n) = 1) \leq \alpha. \tag{5.16}$$

Hence, a test that satisfies (5.16) is a valid NHST test at significance level α, independently of the stopping rule used. If a researcher can obtain a maximum of n observations, the probability of incorrectly selecting the complex model will remain bounded away from one, regardless of the actual number of observations made.

It is well-known that Bayes factor model selection provides a robust null hypothesis test with $t(\alpha) = \alpha$ for *all* fixed α between 0 and 1, as long as $\mathcal{M}_0$ is a singleton. In other words, we may view the output of BFMS as a 'robust' variation of the p-value. This was already noted by Edwards et al. (1963) and interpreted as a frequentist justification for BFMS; it also follows immediately from the following result.

Theorem 5.6 (Special Case of Eq. (2) of Shafer et al. (2011))**.** Let $\mathcal{M}_0, \mathcal{M}_1, M_0$ and M_1 be as in Theorem 5.5 with common support $\mathcal{X} \subset \mathbb{R}^d$ for some $d > 0$. Let $(X_1, X_2, \ldots)$ be an infinite sequence of random vectors all with support $\mathcal{X}$, and fix two distributions, $\bar{\mathbb{P}}_0$ and $\bar{\mathbb{P}}_1$ on $\mathcal{X}^\infty$ (so that under both $\bar{\mathbb{P}}_0$ and $\bar{\mathbb{P}}_1$, $(X_1, X_2, \ldots)$ constitutes a random process). Let, for each n, $\bar{p}_j^{(n)}$ represent the marginal density of $(X_1, \ldots, X_n)$ for the first n outcomes under distribution $\bar{\mathbb{P}}_j$, relative to some product measure ρ^n on $(\mathbb{R}^d)^n$ (we assume $\bar{\mathbb{P}}_0$ and $\bar{\mathbb{P}}_1$ to be such that these densities exist). Then for all $\alpha \geq 0$,

$$\bar{\mathbb{P}}_0\left(\exists n : \frac{\bar{p}_0^{(n)}(X^n)}{\bar{p}_1^{(n)}(X^n)} \leq \alpha\right) \leq \alpha.$$

We first apply this result for Bayes factor model selection, with model priors $\pi_0 = \pi_1 = 1/2$, so that $r_{\text{Bayes}}(x^n) = \mathbb{P}(\mathcal{M}_0|x^n)/\mathbb{P}(\mathcal{M}_1|x^n) = p_{B,0}(x^n)/p_{B,1}(x^n)$. We immediately see that If $M_0 = \{\mu_0\}$ represents a singleton null model, then Bayes factor model selection constitutes a robust hypothesis test for null hypothesis $\mathcal{M}_0$.

What happens if $\mathcal{M}_0$ is not singleton? Full robustness would require that (5.16) holds for all $\mu_0 \in M_0$. The simulations below show that this will in general not be the case for Bayes factor model selection. Yet, the same reasoning as used above implies that we still have some type of robustness in a much weaker sense, which one might call "robustness in prior expectation" relative to prior ω_0 on M_0. Namely, we have for all $0 \leq \alpha \leq 1$:

$$\mathbb{P}_{B,0}(\exists n : \delta_{r,t(\alpha)}(X^n) = 1) \leq \alpha, \tag{5.17}$$

where $\mathbb{P}_{B,0}$ is the Bayes marginal distribution under prior ω_0. In other words, if the beliefs of a Bayesian who adopts prior ω_0 on model $\mathcal{M}_0$ were accurate, then the BFMS method would still give robust p-values, independently of the stopping rule. While for a subjective Bayesian, such a weak form of robustness might perhaps still be acceptable, we will stick to the stronger definition instead, equating 'robust hypothesis tests' with tests satisfying (5.17) uniformly for all $\mu_0 \in M_0$.

5.5.2 AIC is not, and HQ is only asymptotically robust

The situation for AIC is quite different from that for BFMS and switching: for every function $t : (0,1) \to \mathbb{R}_{>0}$, we have, even for every *single* $0 < \alpha < 1$, that $\delta_{AIC,t(\alpha)}$ is *not* a robust null hypothesis test for significance level α. Hence AIC cannot be transformed into a robust test in this sense. This can immediately be seen when comparing a 0-dimensional (fixed mean μ_0) with a 1-dimensional Gaussian location family $\mathcal{M}_1$ (extension to general multivariate exponential families is straightforward but involves tedious manipulations with the Fisher information). Evaluating the left hand side of (5.15) yields that $\delta_{AIC,t(\alpha)}$ will select the complex model if

$$\left| \sum_{i=1}^{n} \tilde{X}_i \right| \geq \frac{\sqrt{2n}}{t(\alpha)}, \tag{5.18}$$

where the $\tilde{X}_i$ are variables with mean 0 and variance 1 if $\mathcal{M}_0$ is correct. Hence, as a consequence of the law of the iterated logarithm (see for example Van der Vaart (1998)), with probability one, infinitely many n exist such that the complex model will be favored, even though it is incorrect.

It is instructive to compare this to the HQ criterion, which, in this example, using the same notation as in (5.18), selects the complex model if

$$\left| \sum_{i=1}^{n} \tilde{X}_i \right| \geq \sqrt{2cn \log \log n}.$$

If $c > 1$ (the case in which HQ is strongly consistent), then this inequality will almost surely not hold for infinitely many n, as again follows from the law of the iterated logarithm. The reasoning can again be extended to other exponential families, and we find that the HQ criterion with $c > 1$ is robust to optional stopping in the crude, asymptotic sense that the probability that there exist infinitely many sample sizes such that the simple model is incorrectly rejected is zero. Yet HQ does not define a robust hypothesis test in the sense above: to get the numerically precise Type I-error bound (5.16) we would need to define $t(\alpha)$ in a model-dependent manner, which is quite complicated in all cases except the Gaussian location families where the asymptotics hold precisely. We note that the same type of asymptotic robustness holds for the BIC criterion as well.

5.5.3 Switching with singleton $\mathcal{M}_0$ is robust under optional stopping

The main insight of this section is simply that, just like BFMS, switching can be used as a robust null hypothesis test as well, as long as $\mathcal{M}_0$ is a singleton: we can view the switch distribution as a model comparison method that outputs odds ratio $r_{\mathrm{sw}}(x^n) = p_{B,0}(x^n)/p_{\mathrm{sw},1}(x^n)$. Until now, we used it to select model 1 if $r_{\mathrm{sw}}(x^n) \leq 1$. If instead we fix a significance level α and select model 1 if $r_{\mathrm{sw}}(x^n) \leq \alpha$, then we immediately see, by applying Theorem 5.6 in the same way as for the Bayes factor case, that r_{sw} constitutes a robust null hypothesis test as long as $\mathcal{M}_0$ is a singleton model. Similarly – at least if the priors involved in the switch criterion are chosen independently of the stopping rule —

just like BFMS, the result $r_{\text{sw}}(x^n)$ of model comparison by switching does not depend on the 'sampling intentions' of the analyst, thus addressing the two most problematic issues with Neyman-Pearson testing. Yet, from a frequentist perspective, switching is preferable to BFMS, since it has substantially better power (type-II error) properties. As could already be seen from Yang's decomposition (5.12), there is an intimate connection between Type-II error and the risk rate achieved by any model comparison method. Formally, we have the following result, a direct corollary of Theorem 5.14 of Section 5.7, which is itself a major building block of our main result Theorem 5.5 (plug in $\gamma = \alpha^{-1}$ into (5.46) to get the corollary):

Corollary 5.7. Using the same notations and under the same conditions as Theorem 5.5, for any $\alpha > 0$, there exist constants $C_1, C_2 > 0$ such that, for every CINECSI subset M_1' of M_1, for every sequence $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ of elements of M_1' with for all n, $\inf_{\mu_0 \in M_0} \|\mu_1^{(n)} - \mu_0\|_2^2 \geq C_1(\log\log n)/n$, we have

$$\mathbb{P}_{\mu_1^{(n)}}(r_{\text{sw}}(x^n) \geq \alpha) \leq \frac{C_2}{\log n}. \tag{5.19}$$

Hence, for any fixed significance level, the power of testing by switching goes to 1 as long as the data are sampled from a distribution $\mu_1^{(n)}$ in M_1 that is farther away from M_0 than order $(\log\log n)/n$; for BFMS, the power only goes to 1 if $\mu^{(n)}$ is farther away than order $O((\log n)/n)$.

Robustness to optional stopping (and hence 'almost the best of three worlds') only holds if $\mathcal{M}_0$ is a singleton; if $\mathcal{M}_0$ is composite, then — using again the same argument as for the Bayes factor case — we immediately see from Theorem 5.6 that the much weaker 'prior expected robustness' property (5.17) still holds. But, the simulations below show that full robustness does fail if μ_0 is 'atypical', i.e. if it resides far out in the tails of the prior ω_0. A major question for future work is now obviously whether there exist versions of the switch criterion that give a truly robust null hypothesis test even under a composite null hypothesis $\mathcal{M}_0$. We return to this question in Section 5.6.

5.5.4 Simulation study

We now provide a simulation to illustrate the differences between AIC, BIC, HQ and the switch criterion in terms of consistency, strong consistency and robustness to optional stopping, illustrating the insights of the previous subsections. In each setting, two of the following three models are compared:

- $\mathcal{M}_0 = \{\mathcal{N}(0,1)\}$.
- $\mathcal{M}_1 = \{\mathcal{N}(\mu,1), \mu \in \mathbb{R}\}$, with a normal prior with mean zero and variance equal to 100 on μ.
- $\mathcal{M}_2 = \{\mathcal{N}(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}\}$, with a normal-inverse-gamma prior: $\mu|\sigma^2 \sim \mathcal{N}(0, C\sigma^2), \sigma^2 \sim IG(\alpha,\beta)$, with $C = 100, \alpha = 1, \beta = 1$.

To illustrate standard consistency, $\mathcal{M}_1$ and $\mathcal{M}_2$ are considered. In the first setting, $\mathcal{M}_1$ is true. $N = 1000$ data sets of length $n = 2500$ are generated from a standard normal

distribution, and AIC, BIC, HQ with $c = 1.05$ and δ_{sw} are evaluated at each sample size. The average selected model index (0 for $\mathcal{M}_1$, 1 for $\mathcal{M}_2$) is given in Figure 5.1.

In the second setting, $\mathcal{M}_2$ is true. The data is generated from a normal distribution with mean 0 and a variance that is varied. For each value of σ, $N = 1000$ datasets of length $n = 2500$ are generated, and the four model selection criteria are evaluated at that sample size. The average selected model index is given in Figure 5.2.

The results are as expected. When the complex model is true, AIC is most likely to select it, at the cost of inconsistency when the simple model is true. BIC is the slowest to correctly select the complex model and the first to correctly select the simple model. HQ and δ_{sw} show intermediate behaviour, HQ being slightly more likely to select the complex model.

To illustrate strong consistency and optional stopping, three scenarios are considered:

1. $\mathcal{M}_0$ vs $\mathcal{M}_1$, data from a standard normal distribution ("scenario 1" — Theorem 5.6 implies that switching defines a test that is robust with respect to optional stopping).

2. $\mathcal{M}_1$ vs $\mathcal{M}_2$, data from a standard normal distribution ("scenario 2", Theorem 5.6 does not only imply robustness, because null model is composite).

3. $\mathcal{M}_1$ vs $\mathcal{M}_2$, data from a normal distribution with mean 35 and variance 1 ("scenario 3", Theorem 5.6 again does not imply robustness).

We create $N = 1000$ data sets of length $n_{\max} = 10000$ in each scenario. We select the complex model when δ_{sw} is larger than 20 (in terms of the robust p-value interpretation of Theorem 5.6, this corresponds to a significance level of 0.05). We estimate two probabilities at each sample size n:

- The probability that there will ever be a model index after n at which the complex model will be selected (Figure 5.3), approximated by checking whether the complex model is selected at any sample size between n and $3n_{\max}$.

- The probability that there exists a model index before n at which the complex model would have been selected (Figure 5.4).

Figure 5.3 can be interpreted as a check whether strong consistency holds — if it does, then the probabilities should converge to 0 as $n \to \infty$. Van Erven et al.'s (2007) theorem implies that strong consistency holds in all three scenarios, and the graphs confirm this — even though for scenario 3, in which data comes from a $\mu \in M_0$ that is 'atypical' under the prior, it takes a bit longer — illustrating that strong consistency is not a uniform notion. The graph also illustrates that strong consistency can be viewed as an asymptotic, nonuniform version of robustness to optional stopping — it implies that from some sample size (which may be very large though) onwards, one will never again falsely reject no matter how long one keeps sampling.

Figure 5.4 refers to nonasymptotic optional stopping: in scenario 1, the conditions from Theorem 5.6 hold, and indeed the figure shows that the probability that the complex model is *ever* incorrectly selected even when optional stopping is used, is bounded by 0.05 (the observed bound is 0.015). In scenarios 2 and 3, the conditions from Theorem 5.6 do not hold. In scenario 2, the behaviour of the switch criterion is similar to scenario 1. However,

in scenario 3, the probability of a false rejection opportunity before sample size n is not bounded by 0.05, but quickly goes to 0.15. We clearly see that δ_{sw} is not robust to optional stopping in scenario 3.

When the simplest model is not a singleton, the choice of prior on the model parameters (in scenarios 2 and 3 on μ in $\mathcal{M}_1$ and on (μ, σ^2) in $\mathcal{M}_2$) affects the results. In both scenario 2 and 3, δ_{sw} must still satisfy the weak, prior-expected version of robustness (5.17), as we have seen in Section 5.5.3. In scenario 2, the prior is centered at the data-generating value of zero and we do observe actual robustness. In scenario 3 however, the prior is centered at zero while the data is generated with a mean of 35, 3.5 standard deviations away from the prior mean — thus μ is 'atypical' under the prior, and, as the figure shows, nonasymptotic robustness is violated.

5.6 Discussion and future work

In this paper we showed that switching combines near-rate optimality, consistency and, for singleton $\mathcal{M}_0$, robustness to optional stopping. We end the paper by highlighting three issues which, we feel, need additional discussion: first, the desirability of consistency; second, whether there is anything 'special' to the switch criterion as opposed to other possible trade-offs between risk optimality and consistency; and third, the limitations of switching in its current form.

Consistency Since the desirability of consistency, in the sense of finding the smallest model containing the true distribution, is somewhat controversial, let us discuss it a bit further. The main argument against consistency is made by those adhering to Box's maxim 'Essentially, all models are wrong, but some are useful' (Box and Draper, 1987). According to some, the goal of model selection should therefore not be to select a non-existing 'true' model, but to obtain the best predictive inference or best inference about a parameter (Burnham and Anderson, 2004; Forster, 2000). Another issue with consistency is that it is a 'nonuniform' notion, which in our context means that — as is indeed easy to see — it is impossible to give a bound on the probability under $\mathbb{P}_\mu$ of selecting the wrong model at sample size n that converges to 0 uniformly for all $\mu \in M$. This nonuniformity implies that consistency is of little practical consequence for post-model selection inference (Leeb and Pötscher, 2005).

As to the first argument, one can reply that there do exist situations in which a model can be correct, for example in the field of extrasensory perception (Bem, 2011), in which it seems exceedingly likely that the null model (expressing that no such thing exists) is correct; another example is genetic linkage (Gusella et al., 1983; Tsui et al., 1985). The second argument is more convincing, but only to argue that even if consistency holds, a method may not be very useful in practice. It does not contradict that consistency can sometimes be a highly desirable (but never the only highly desirable) property — we feel that this is the case whenever we are not purely interested in prediction but instead are also seeking to find out whether a certain structural relationship (e.g. dependence between variables) holds or not.

Going one step further, it seems a good idea to study model selection methods not in

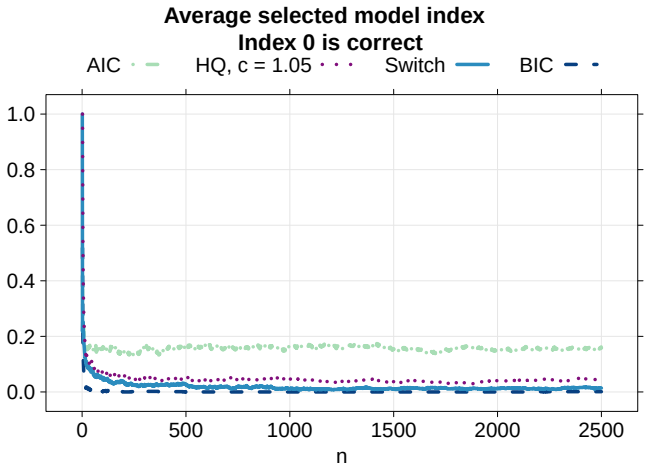

Figure 5.1: $N = 1000$ data sets of length $n = 2500$ are generated from a standard normal distribution and the criteria are evaluated at each sample size. The figure shows the average selected model index (0 for $\mathcal{M}_1$, 1 for $\mathcal{M}_2$). The true index is 0.

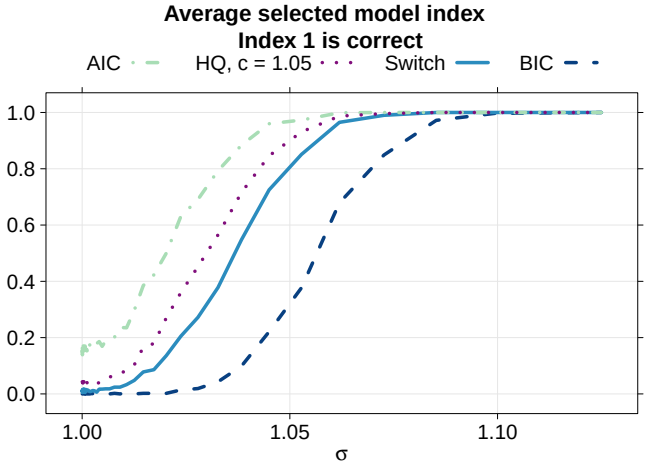

Figure 5.2: $N = 1000$ data sets of length $n = 2500$ are generated from a normal distribution with mean 0 and variance σ^2 for a range of values of σ. The criteria are evaluated at $n = 2500$. The figure shows the average selected model index (0 for $\mathcal{M}_1$, 1 for $\mathcal{M}_2$). The true index is 1.

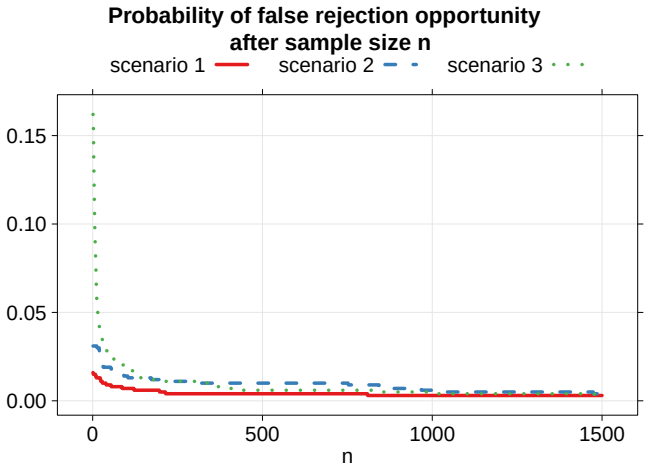

Figure 5.3: $N = 1000$ data sets of length $n_{\max} = 10000$ in each scenario, from the simple model. The complex model is selected when $\delta_{\text{sw}}(x^n) > 20$. Estimated probability that there exists a model index after n at which the complex model will be selected. Results shown up to $n = 1500$ for clarity. After $n = 1500$, the three curves are indistinguishable and all very close to zero.

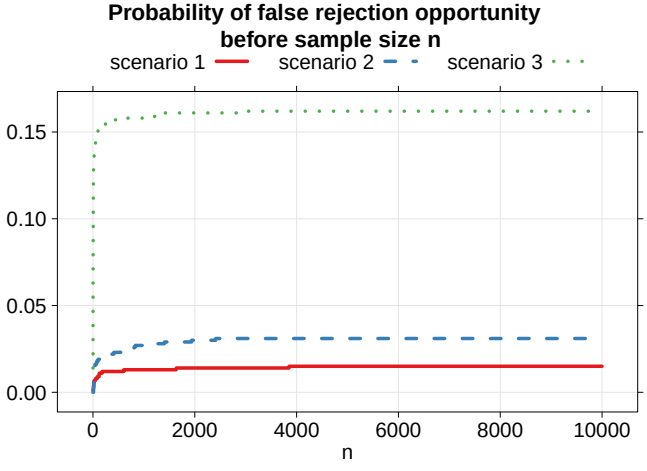

Figure 5.4: Setting as Figure 5.3. Estimated probability that there exists a model index before n at which the complex model would have been selected.

terms of the asymptotic, nonuniform notion of consistency but instead by a more tangible finite-sample analogue. For the case of just two models, Type-I and Type-II errors provide exactly this analogue — note that if both errors go to 0 as $n \to \infty$, this implies consistency. Thus, the *practical* importance of the present work, for us, is mostly that model comparison by switching defines, like Bayes, a robust null hypothesis test — providing Type-I errors irrespective of the stopping rule and thus more in line with actual practice — yet has better Type-II error behaviour, allowing the Type-II error to become small (i.e. the power to go to 1) whenever the true distribution sits at a distance of order $\sqrt{(\log \log n)/n}$ rather than $\sqrt{(\log n)/n}$, as with Bayes. We only showed robustness for singleton $\mathcal{M}_0$, however, and our simulations show that it may fail for composite $\mathcal{M}_0$, so *the* major goal for future work is therefore, to come up with methods that are robust to optional stopping also under composite $\mathcal{M}_0$.

How special is the switch distribution? Since Yang proved that in general, the conflict between consistency and risk-optimality is not resolvable, one might argue that any model selection rule just picks some position in the spectrum of behaviours of consistency vs. risk-optimality. For example, one might have a modified HQ criterion which picks $\mathcal{M}_1$ if, using the same setup and notation as in (5.18),

$$\left| \sum_{i=1}^{n} \tilde{X}_i \right| \geq \sqrt{n \log \log \log n}. \tag{5.20}$$

By the central limit theorem, such a method will be consistent, yet when combined with an efficient estimator will achieve the minimax estimation rate up to a $\log \log \log n$ factor, improving on the switch criterion by an additional logarithm. Note however that both the switch distribution and HQ (with $c > 1$) achieve *strong* consistency. The meaning of strong consistency is illustrated in Figure 5.3 above: it means that, from some n onward, the wrong model will never be selected any more, no matter how long one keeps sampling. It is easy to see from the law of the iterated logarithm that any strongly consistent method can have rate no faster than order $(\log \log n)/n$ — in particular, (5.20) is not strongly consistent. Thus, in this sense both switching and HQ do take a special place in the consistency vs. risk-optimality spectrum as obtaining the fastest rates compatible with strong consistency, which may be viewed as asymptotic robustness to optional stopping. While this may mostly be of theoretical interest, the switch distribution also takes a special place in terms of its nonasymptotic robustness to optional stopping: again, the law of the iterated logarithm implies that any model comparison method that defines a robust hypothesis test cannot achieve estimation rate better than order $(\log \log n)/n$. Again, the main open question here is whether one can modify it so that robustness for composite $\mathcal{M}_0$ is achieved as well.

Future work — limitations of the switch distribution and our results Whereas the results in this paper all apply to the original switch distribution as defined by Van Erven et al. (2007) and a simplification thereof, for full robustness to optional stopping with composite $\mathcal{M}_0$, some substantial changes have to be made, as suggested by the results in Figure 5.4. Initial research suggests that such a modification of the switch distribution

might indeed be constructed, based on techniques in Ramdas and Balsubramani (2015); whereas, compared to Bayes factor testing, in the current switch criterion, $p_{B,1}$ is modified to another distribution and $p_{B,0}$ can remain the same, in this new version we would also have to change $p_{B,0}$ — the resulting distribution would not have a Bayesian interpretation any more. While this work is still under development, to avoid the nonrobustness seen in Figure 5.4 as much as possible, for the time being we recommend using flat priors (but in this case, not completely flat - Jeffreys' prior on μ is improper, in which case Theorem 5.6 holds in none of the scenarios and simulations — not reported here — show that optional stopping robustness is violated).

Another limitation lies not in the switch distribution, but in our results: these are restricted to two nested exponential family models. It would be interesting to extend them to more than two models — highlighting the distinction between model selection and testing — and going beyond exponential families. We are hopeful that switching still behaves well in such contexts — we note that the risk rate convergence results of Van Erven et al. (2012) were for countable, possibly infinite collections of completely general models — but they invariably dealt with the cumulative risk. While all our experiments suggest that small cumulative risk usually goes together with small instantaneous risk, formal analysis of the switch criterion's instantaneous risk is far more difficult, and the present paper heavily relies on sufficiency to do so — so extension of our results beyond exponential families would be difficult.

Before doing so, we would prefer to modify the switch distribution further, since the present version has a drawback when used in nonsequential settings: the precise results it gives are dependent on the order of the data, even if all the models under consideration are i.i.d. Thus, it would be interesting and challenging to design an alternative, order-independent method that, like the switch distribution, is strongly consistent, near rate- and power-optimal, and is robust to optional stopping under composite $\mathcal{M}_0$. Such a method would essentially truly achieve the best of the three worlds we considered in this paper — and this is the method we aim for in our future research.

Acknowledgements

The central result of this paper, Theorem 5.5, already appeared in the Master's Thesis (Van der Pas, 2013) for the special case where $m_1 = 1$ and $m_0 = 0$, but the proof supplied there contained an error. We are grateful to Tim van Erven for pointing this out to us.

5.7 Proofs

In this appendix, we start by listing some well-known properties of exponential families which we will repeatedly use in the proofs. Then, in Section 5.7.4, we provide a sequence of technical lemmata that lead up to the proof of our main result, Theorem 5.5. Finally, in Section 5.7.5, we compare the switch distribution and criterion as defined here to the original switch distribution and criterion of Van Erven et al. (2012).

Additional notation Our results will often involve displays involving several constants. The following abbreviation proves useful: when we write 'for positive constants $\vec{c}$, we have ...', we mean that there exist some $(c_1, \ldots, c_N) \in \mathbb{R}^N$, with $c_1, \ldots, c_N > 0$, such that ... holds; here N is left unspecified but it will always be clear from the application what N is. Further, for positive constants $\vec{b} = (b_1, b_2, b_3)$, we define $\texttt{small}_{\vec{b}}(n)$ as

$$\texttt{small}_{\vec{b}}(n) = \begin{cases} 1 & \text{if } n < b_1 \\ b_2 e^{-b_3 n} & \text{if } n \geq b_1, \end{cases}$$

and we frequently use the following fact. Suppose that $\mathcal{E}_1, \mathcal{E}_2, \ldots$ is a sequence of events such that $\mathbb{P}(\mathcal{E}_n) \leq \texttt{small}_{\vec{b}}(n)$. Then we also have, for any event $\mathcal{A}$, and for all n,

$$\mathbb{P}(\mathcal{A}, \mathcal{E}_n^c) \geq \mathbb{P}(\mathcal{A}) - \texttt{small}_{\vec{b}}(n), \tag{5.21}$$

as is immediate from $\mathbb{P}(\mathcal{A}, \mathcal{E}_n^c) = \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{A}, \mathcal{E}_n) \geq \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{E}_n)$.

The components of a vector $\mu \in \mathbb{R}^n$ are given by $(\mu_1, \mu_2, \ldots, \mu_n)$. If the vector already has an index, we add a comma, for example $\mu_1 = (\mu_{1,1}, \mu_{1,2}, \ldots, \mu_{1,n})$. A sequence of vectors is denoted by $\mu^{(1)}, \mu^{(2)}, \ldots$.

5.7.1 Definitions concerning and properties of exponential families

The following definitions and properties can all be found in the standard reference (Barndorff-Nielsen, 1978) and, less formally, in (Grünwald, 2007, Chapters 18 and 19).

A k-dimensional exponential family is a set of distributions on $\mathcal{X}$, which we invariably represent by the corresponding set of densities $\{p_\theta \mid \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$, such that any member p_θ can be written as

$$p_\theta(x) = \frac{1}{z(\theta)} e^{\theta^T \varphi(x)} r(x) = e^{\theta^T \varphi(x) - \psi(\theta)} r(x), \tag{5.22}$$

where $\varphi(x) = (\varphi_1(x), \ldots, \varphi_k(x))$ is a *sufficient statistic*, r is a non-negative function called the *carrier*, z the *partition function* and $\psi(\theta) = \log z(\theta)$. We assume the representation (5.7.1) to be *minimal*, meaning that the components of $\varphi(x)$ are linearly independent.

The parameterization in (5.22) is referred to as the *canonical* or *natural parameterization*; we only consider families for which the set Θ is open and connected. Every exponential family can alternatively be parameterized in terms of its *mean-value parameterization*, where the family is parameterized by the mean $\mu = \mathbb{E}_\theta[\varphi(X)]$, with μ taking values in $M \subset \mathbb{R}$, where μ as a function of θ is smooth and strictly increasing; as a consequence, the set M of mean-value parameters corresponding to an open and connected set Θ is itself also open and connected. Whenever for data $x_1, \ldots, x_n$, we have $\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \in M$, then the maximum likelihood is uniquely achieved by the μ that is itself equal to this value,

$$\widehat{\mu}(x^n) = \frac{1}{n} \sum_{i=1}^n \varphi(x_i). \tag{5.23}$$

We thus define the maximum likelihood estimator (MLE) to be equal to (5.23) whenever $\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \in M$. Since the result below which directly involves the MLE (Lemma 5.11)

does not depend on its value for x^n with $\frac{1}{n}\sum_{i=1}^n \varphi(x_i) \notin M$, we can leave $\widehat{\mu}(x^n)$ undefined for such values. However, if we want to use the MLE as a 'sufficiently efficient' estimator as used in the statement of Theorem 5.5, we need to define $\widehat{\mu}(x^n)$ for such values in such a way that (5.13) is satisfied, as illustrated in Example 1.

A standard property of exponential families says that, for any $\mu \in M$, any distribution $\mathbb{Q}$ on $\mathcal{X}$ with $\mathbb{E}_{X\sim\mathbb{Q}}[\varphi(X)] = \mu$, any $\mu' \in M$, we have

$$\mathbb{E}_{X\sim\mathbb{Q}}\left[\log \frac{p_\mu(X)}{p_{\mu'}(X)}\right] = \mathbb{E}_{X\sim\mathbb{P}_\mu}\left[\log \frac{p_\mu(X)}{p_{\mu'}(X)}\right] = D(\mu\|\mu'), \tag{5.24}$$

the final equality being just the definition of $D(\cdot\|\cdot)$. Now fix an arbitry sample x^n. By taking $\mathbb{Q}$ to be the empirical distribution on $\mathcal{X}$ corresponding to sample x^n, it follows from (5.24) that if $\widehat{\mu}(x^n) \in M$ then also the following relationship holds for any $\mu' \in M$:

$$\frac{1}{n}\log \frac{p_{\widehat{\mu}(x^n)}(x^n)}{p_{\mu'}(x^n)} = D(\widehat{\mu}(x^n)\|\mu'). \tag{5.25}$$

(5.24) and (5.25) are a direct consequence of the sufficiency of $\widehat{\mu}_1(X^n)$, and folklore among information theorists. For a proof of (5.24) and more details on (5.25), see e.g. (Grünwald, 2007, Chapter 19), who calls this the *robustness property* of the KL divergence for exponential families.

We are now in a position to prove Proposition 5.2, which we repeat for convenience.

Proposition 5.2 Let M, a product of open intervals, be the mean-value parameter space of an exponential family, and let M' be an CINECSI subset of M. Then there exist positive constants $\vec{c}$ such that for all $\mu, \mu' \in M'$,

$$c_1\|\mu' - \mu\|_2^2 \le c_2 \cdot d_{ST}(\mu'\|\mu) \le d_{H^2}(\mu', \mu) \le d_R(\mu', \mu) \le D(\mu'\|\mu) \le c_3\|\mu' - \mu\|_2^2. \tag{5.26}$$

and for all $\mu' \in M', \mu \in M$ (i.e. μ is now not restricted to lie in M'),

$$d_{H^2}(\mu', \mu) \le c_4\|\mu' - \mu\|_2^2 \le c_5 \cdot d_{ST}(\mu'\|\mu) \le c_6\|\mu' - \mu\|_2^2. \tag{5.27}$$

Proof. We start with (5.26). The third and fourth inequality are immediate by using $-\log x \ge 1 - x$ and Jensen's inequality, respectively. From standard properties of Fisher information for exponential families (Barndorff-Nielsen, 1978) we have that, for any CINECSI (hence compact and bounded away from the boundaries of M) subset M' of M, there exists positive $\vec{C}$ with

$$0 < C_1 = \inf_{\mu\in M'} \det I(\mu) < \sup_{\mu\in M'} \det I(\mu) = C_2 < \infty, \tag{5.28}$$

from which we infer that for all $\mu' \in M'$, $\mu, \mu'' \in \mathbb{R}^m$,

$$C_3\|\mu - \mu''\|_2^2 \le (\mu - \mu'')^T I(\mu')(\mu - \mu'') \le C_4\|\mu - \mu''\|_2^2, \tag{5.29}$$

for some $0 < C_3 \le C_4 < \infty$. Using (5.29), the first inequality is immediate, and the final inequality follows straightforwardly from a second-order Taylor approximation of KL divergence as in (Grünwald, 2007, Chapter 4). It only remains to establish the second

inequality. Now, since M' is CINECSI and hence compact the fifth (rightmost) inequality implies that there is a $C_5 < \infty$ such that $\sup_{\mu,\mu' \in M'} D(\mu' \| \mu) < C_5$ and hence, via the fourth inequality, that $\sup_{\mu,\mu' \in M'} d_R(\mu', \mu) < C_5$. Equality (5.5) now implies that there is a C_6 such that

$$\sup_{\mu,\mu' \in M'} d_R(\mu', \mu)/d_{H^2}(\mu', \mu) < C_6. \tag{5.30}$$

Using again (5.28), a second order Taylor approximation as in Van Erven and Harremoës (2014) now gives that for some constant $C_7 > 0$, $\|\mu - \mu'\|_2^2 \leq C_7 d_R(\mu', \mu)$ for all $\mu, \mu' \in M'$. The first result, (5.26), now follows upon combining this with (5.30).

As to (5.27), the second and third inequality are immediate from (5.29). For the first inequality, note that, since M' is CINECSI and we assume M to be a product of open intervals, there must exist another CINECSI subset M'' of M strictly containing M' such that $\inf_{\mu' \in M', \mu \in M \setminus M''} \|\mu' - \mu\|_2^2 = \delta$ for some $\delta > 0$. We now distinguish between μ in (5.27) being an element of (a) M'' or (b) $M \setminus M''$. For case (a) (5.26), with M'' in the role of M', gives that there is a constant C_8 such that for all $\mu \in M''$, $d_{H^2}(\mu', \mu) \leq C_8 \|\mu' - \mu\|_2^2$. For case (b), $\mu \in M \setminus M''$, we have $\|\mu' - \mu\|_2^2 \geq \delta$ and, using that squared Hellinger distance for any pair of distributions is bounded by 2, we have $d_{H^2}(\mu', \mu) \leq (2/\delta)\|\mu' - \mu\|_2^2$. Thus, by taking $c_4 = \max\{C_8, 2/\delta\}$, case (a) and (b) together establish the first inequality in (5.27). □

5.7.2 Preparation for proof of main result: results on large deviations

Let $\mathcal{M}_1$ and M_1 be as in Theorem 5.5. For the following result, Lemma 5.8, we set $\widehat{\mu}_1'(X^n) := n^{-1} \sum \varphi(X_i)$, so that $\widehat{\mu}_1'(X^n) = \widehat{\mu}_1(X^n)$ whenever $n^{-1} \sum \varphi(X_i) \in M_1$. It is essentially a multidimensional extension of a standard information-theoretic result, with KL divergence replaced by squared error loss. The result states the following: whenever $\mathcal{M}_1$ is a single-parameter exponential family (that is, $m_1 = 1$), then for any $\mu \in M_1$, all $a, a' > 0$ with $\mu + a \in M_1$, $\mu - a' \in M_1$,

$$\mathbb{P}_\mu(\widehat{\mu}_1'(X^n) \geq \mu + a) \leq e^{-nD(\mu+a\|\mu)}. \;;\; \mathbb{P}_\mu(\widehat{\mu}_1'(X^n) \leq \mu - a') \leq e^{-nD(\mu-a'\|\mu)}. \tag{5.31}$$

For a simple proof, see (Grünwald, 2007, Section 19.4.2); for discussion see (Csiszár, 1984) — the latter reference gives a multidimensional extension of (5.31) but of a very different kind than Lemma 5.8 below. To prepare for the lemma, let $\mathcal{M}_1$ and M_1 be as in Theorem 5.5 and, for any $\mu \in M_1$ and any $\vec{a}, \vec{b} \in \mathbb{R}_{>0}^{m_1}$, define the ℓ_∞-rectangle $R_\infty(\mu, \vec{a}, \vec{b}) = \{\mu' \in \mathbb{R}^{m_1} : \forall j = 1, \ldots, m_1, -b_j \leq \mu_j' - \mu_j \leq a_j\}$.

Lemma 5.8. Let $\mathcal{M}_1$ and M_1 be as in Theorem 5.5 and fix an arbitrary CINECSI subset M_1' of M_1. Then there is a $c > 0$ (depending on M_1') such that, for all $\mu \in M_1$, all n, all $\vec{a}, \vec{b} \in \mathbb{R}_{>0}^{m_1}$ such that $R_\infty(\mu, \vec{a}, \vec{b}) \subset M_1'$,

$$\mathbb{P}_\mu(\widehat{\mu}_1'(X^n) \notin R_\infty(\mu, \vec{a}, \vec{b})) \leq 2m_1 e^{-nc \cdot (\min_j \min\{a_j, b_j\})^2}. \tag{5.32}$$

Proof. For $j = 1, \ldots, m_1$, $d \in \mathbb{R}$, let $\vec{e_j}$ represent the jth standard basis vector, such that $\mu + d\vec{e_j} = (\mu_1, \ldots, \mu_{j-1}, \mu_j + d, \mu_{j+1}, \ldots, \mu_{m_1})$, and let $D_{\mu + d\vec{e_j}} := D(\mu + d\vec{e_j} \| \mu)$. We now have that

there exist constants $c_{a,1},\ldots,c_{a,m_1},c_{b,1},\ldots,c_{b,m_1} > 0$ such that for $c := \min\{c_{a,1},\ldots,c_{a,m_1},c_{b,1},\ldots,c_{b,m_1}\}$, all n,

$$\begin{aligned}\mathbb{P}_\mu(\widehat{\mu}_1(X^n) \notin R_\infty(\mu,\vec{a},\vec{b})) &\leq \sum_{j=1}^{m_1} \mathbb{P}_\mu(\widehat{\mu}_{1,j}(X_n) \geq \mu_j + a_j) + \sum_{j=1}^{m_1} \mathbb{P}_\mu(\widehat{\mu}_{1,j}(X^n) \leq \mu_j - b_j) \\ &\leq \sum_{j=1}^{m_1} \left(e^{-nD_{\mu+a_j\vec{e}_j}} + e^{-nD_{\mu-b_j\vec{e}_j}}\right) \leq \sum_{j=1}^{m_1} \left(e^{-nc_{a,j}a_j^2} + e^{-nc_{b,j}b_j^2}\right) \\ &\leq 2m_1 e^{-nc\cdot(\min_j \min\{a_j,b_j\})^2},\end{aligned}$$

Here the first inequality follows from the union bound, and the second follows by applying, for each of the $2m_1$ terms, (5.31) above to the one-dimensional exponential sub-family $\{p_\mu \mid \mu \in M_1 \cap \{\mu : \mu = \mu + d\vec{e}_j \text{ for some } d \in \mathbb{R}\}\}$. The third follows by Proposition 5.2 together with the equivalence of the ℓ_2 and sup norms on $\mathbb{R}^{m_1}$, and the final inequality is immediate. □

Lemma 5.9. Under conditions and notations as in Theorem 5.5, let μ,μ' be elements of M_1 and suppose $X^N = (X_{n_1},\ldots,X_{n_2})$ is a sequence of i.i.d. observations of length N from p_μ. Then, for any $A \in \mathbb{R}$:

$$\mathbb{P}_\mu\left(\log\frac{p_\mu(X^N)}{p_{\mu'}(X^N)} < A\right) \leq e^{\frac{1}{2}A}e^{-\frac{N}{2}d_R(\mu',\mu)}. \tag{5.33}$$

Proof. For any A, by Markov's inequality:

$$\begin{aligned}\mathbb{P}_\mu\left(\log\frac{p_\mu(X^N)}{p_{\mu'}(X^N)} < A\right) &= \mathbb{P}_\mu\left(\left(\frac{p_{\mu'}(X^N)}{p_\mu(X^N)}\right)^{\frac{1}{2}} > e^{-\frac{1}{2}A}\right) \leq e^{\frac{1}{2}A}\mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X^N)}{p_\mu(X^N)}\right)^{\frac{1}{2}}\right] \\ &= e^{\frac{1}{2}A}\left(\mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X_{n_1})}{p_\mu(X_{n_1})}\right)^{\frac{1}{2}}\right]\right)^N = e^{\frac{1}{2}A}e^{\log\left(\mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X_{n_1})}{p_\mu(X_{n_1})}\right)^{\frac{1}{2}}\right]\right)^N} \\ &= e^{\frac{1}{2}A}e^{-\frac{N}{2}\left(-\frac{1}{1-1/2}\log\mathbb{E}_\mu\left[\left(\frac{p_{\mu'}(X_{n_1})}{p_\mu(X_{n_1})}\right)^{\frac{1}{2}}\right]\right)} = e^{\frac{1}{2}A}e^{-\frac{N}{2}d_R(\mu,\mu')}.\end{aligned} \tag{5.34}$$

□

Proposition 5.10. Let $\mathcal{M}_0,\mathcal{M}_1,M_0,M_1$ be as in Theorem 5.5 and let M_1' be a CINECSI subset of M_1. Then there exists another, larger, CINECSI subset M_1'' of M_1 and positive constants $\vec{b}$ such that M_1' is itself a CINECSI subset of M_1'' and for both $j \in \{0,1\}$, the ML estimator $\widehat{\mu}_j(x^n)$ satisfies

$$\sup_{\mu\in M_1'} P_\mu(\widehat{\mu}_j(X^n) \notin M_1'') \leq \mathtt{small}_{\vec{b}}(n).$$

Proof. M_1 can be written as in (5.10), and hence we can define a set

$$M_1'' = [\zeta_{1,1}^*, \eta_{1,1}^*] \times \ldots \times [\zeta_{1,m_1}^*, \eta_{1,m_1}^*]$$

for values $\zeta_{1,j}^*, \eta_{1,j}^* \in \mathbb{R}$ such that M_1'' is a CINECSI subset of M_1. Since M_1' is connected with compact closure in interior of M_1 and M_1'' is a subset of M_1, we can choose the $\zeta_{1,j}^*, \eta_{1,j}^* \in \mathbb{R}$ such that M_1' is itself a CINECSI subset of M_1''. Since M_1' is connected and its closure is in the interior of M_1'' which is itself compact, it follows that there is some $\delta > 0$ such that, for all $\mu_1' \in M_1', \mu_1'' \notin M_1''$, all $j \in \{1, \ldots, m_1\}$, it holds $|\mu_{1,j}' - \mu_{1,j}''| > \delta$. It now follows from Lemma 5.8, applied with $\vec{a}$ chosen such that $R_\infty(\mu', \vec{a}) = M_1''$, that for every $\mu' \in M_1'$, all n,

$$\mathbb{P}_{\mu'}\left(\widehat{\mu}_1(X^n) \notin M_1''\right) \leq C_1 e^{-nC_2\delta^2}$$

for some constants C_1, C_2. Here we used that by construction, each entry of $\vec{a}$ must be at least as large as δ. Since $\widehat{\mu}_{1,j}(x^n)$ and $\widehat{\mu}_{0,j}(x^n)$ coincide for $0 < j \leq m_0$ and $\widehat{\mu}_{0,j}(x^n)$ is constant for $m_0 < j \leq m_1$, the result follows for $\widehat{\mu}_0(x^n)$ as well. □

5.7.3 Preparation for proof of main result: results on Bayes factor model selection

Lemma 5.11. Let $\mathcal{M}_0, \mathcal{M}_1, M_0, M_1$ be as in Theorem 5.5 and let, for $j \in \{0,1\}$, M_j' be a CINECSI subset of M_j. For both $j \in \{0,1\}$, there exist positive constants $\vec{c}, \vec{b}$ such that for all $\mu_1 \in M_1'$,

$$c_1 \leq n^{-m_j/2} \cdot \frac{p_{\widehat{\mu}_j(X^n)}(X^n)}{p_{B,j}(X^n)} \leq c_2, \tag{5.35}$$

with $\mathbb{P}_{\mu_1}$-probability at least $1 - \texttt{small}_{\vec{b}}(n)$.

Proof. For a Bayesian marginal distribution p_B defined relative to m-dimensional exponential family $\mathcal{M}$ given in its mean-value parameterization M, with a prior $\omega(\cdot)$ that is continuous and strictly positive on M, we have as a consequence of the familiar Laplace approximation of the Bayesian marginal distribution of exponential famlies as in e.g. (Kass and Raftery, 1995),

$$p_B(x^n) \sim \left(\frac{n}{2\pi}\right)^{-m/2} \cdot \frac{\omega(\widehat{\mu}(x^n))}{\sqrt{\det I(\widehat{\mu}(x^n))}} p_{\widehat{\mu}(x^n)}(x^n).$$

As shown in Theorem 8.1 in (Grünwald, 2007), this statement holds uniformly for all sequences x^n with ML estimators in any fixed *CINECSI* subset M' of M. By compactness of M', and by positive definiteness and continuity of Fisher information for exponential families, the quantity $\omega(\widehat{\mu})/\sqrt{\det I(\widehat{\mu})}$ will be bounded away from zero and infinity on such sequences, and, applying the result to both the families $\mathcal{M}_0$ and $\mathcal{M}_1$ it follows that there exist $c_1, c_2 > 0$ such that for all n larger than some n_0, uniformly for all sequences x^n with $\widehat{\mu}_j(x^n) \in M_j'$, we have:

$$c_1 \leq n^{-m_j/2} \cdot \frac{p_{\widehat{\mu}_j(x^n)}(x^n)}{p_{B,j}(x^n)} \leq c_2. \tag{5.36}$$

The result now follows by combining this statement with Proposition 5.10. □

Lemma 5.12. Let $\mathcal{M}_0, \mathcal{M}_1, M_0, M_1$ and the Bayesian marginal distribution $p_{B,0}$ be as in Theorem 5.5. Let M_1' be a CINECSI subset of M_1. Then there exist positive constants $\vec{c}$ and $\vec{b}$ such that for all n, all $\mu_1 \in M_1'$, all $A \in \mathbb{R}$,

$$\mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A\right) \leq n^{m_1/2} \cdot c_1 \cdot e^{\frac{1}{2}c_2 A} e^{-\frac{n}{2}c_3 \|\mu_1 - \mu_0\|_2^2} + \texttt{small}_{\vec{b}}(n),$$

where for each μ_1, $\mu_0 = \Pi_0(\mu_1)$ as in (5.11).

Proof. Fix constants C_1, C_2 such that they are smaller and larger respectively than the constants c_1, c_2 from Lemma 5.11 and define

$$\mathcal{E}_n = \left\{X^n : C_1 \leq n^{-m_1/2} \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{B,1}(X^n)} \leq C_2\right\}.$$

Using Lemma 5.11, we have that there exists positive $\vec{b}$ such that for all $A \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A\right) \\
=&\mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A, \mathcal{E}_n\right) + \mathbb{P}_{\mu_1}\left(\log \frac{p_{B,1}(X^n)}{p_{B,0}(X^n)} < A, \mathcal{E}_n^c\right) \\
\leq&\mathbb{P}_{\mu_1}\left(\log \frac{C_2^{-1} n^{-m_1/2} p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{B,0}(X^n)} < A, \mathcal{E}_n\right) + \texttt{small}_{\vec{b}}(n) \\
\leq&\mathbb{P}_{\mu_1}\left(\log \frac{C_2^{-1} n^{-m_1/2} p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A\right) + \texttt{small}_{\vec{b}}(n) \\
=&\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A + \log C_2 n^{m_1/2}\right) + \texttt{small}_{\vec{b}}(n). \qquad (5.37)
\end{aligned}$$

To bound this probability further, we need to relate $p_{B,0}$ to $p_{B',0}$, the Bayesian marginal likelihood under model M_0 under a prior with support restricted to a compact set M_0'. To define M_0', note first that there must exist a CINECSI subset, say M_1'', of M_1 such that M_1' is itself a CINECSI subset of M_1''. Take any such M_1'' and let M_0' be the closure of $M_1'' \cap M_0$. Given ω, the prior density on $\Pi'(M_0)$ used in the definition of $p_{B,0}$, define $\omega'(\nu) = \omega(\nu)/\int_{\nu \in \Pi'(M_0')} \omega(\nu) d\nu$ as the prior density restricted to and normalized on $\Pi'(M_0')$ and let $p_{B',0}$ be the corresponding Bayesian marginal density on X^n.

To continue bounding (5.37), define

$$\mathcal{E}_n' = \left\{X^n : C_3 \leq n^{-m_0/2} \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{B,0}(X^n)} \leq C_4 \text{ and } C_3 \leq n^{-m_0/2} \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{B',0}(X^n)} \leq C_4\right\},$$

with C_3 and C_4 smaller and larger respectively than the constants c_1 and c_2 resulting from Lemma 5.11 (note that Lemma 5.11 can be applied to $p_{B',0}$ as well, by taking M_0 in that lemma to be the interior of M_0' as defined here). Set $C_5 > C_4/C_3$, and note that for any $A_1 \in \mathbb{R}$, abbreviating $\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{C_5 p_{B',0}(X^n)} < A_1\right)$ to p^*, we have

$$\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A_1\right)$$

$$
\begin{aligned}
=&\mathbb{P}_{\mu_1}\left(\log\frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A_1,\ \frac{p_{B_0}(X^n)}{p_{B',0}(X^n)} < C_5\right) + \mathbb{P}_{\mu_1}\left(\log\frac{p_{\mu_1}(X^n)}{p_{B,0}(X^n)} < A_1,\ \frac{p_{B_0}(X^n)}{p_{B',0}(X^n)} \geq C_5\right)\\
\leq&\mathbb{P}_{\mu_1}\left(\log\frac{p_{\mu_1}(X^n)}{C_5 p_{B',0}(X^n)} < A_1\right) + \mathbb{P}_{\mu_1}\left(p_{B,0}(X^n) \geq C_5 p_{B',0}(X^n)\right)\\
=&p^* + \mathbb{P}_{\mu_1}\left(p_{B,0}(X^n) \geq C_5 p_{B',0}(X^n)\right)\\
\leq&p^* + \mathbb{P}_{\mu_1}\left(\frac{p_{B,0}(X^n)}{p_{B',0}(X^n)} \geq C_5, \mathcal{E}'_n\right) + \mathbb{P}_{\mu_1}\left(\frac{p_{B,0}(X^n)}{p_{B',0}(X^n)} \geq C_5, (\mathcal{E}'_n)^c\right)\\
\leq&p^* + 0 + \mathtt{small}_{\vec{b}}(n).
\end{aligned}
\tag{5.38}
$$

Now it only remains to bound p^*. To this end, let

$$
C_6 := \int_{\nu\in\Pi'(M'_0)} \sqrt{\omega(\nu)}d\nu. \tag{5.39}
$$

Since M'_0 has compact closure in the interior of M_0 and we are assuming that ω has full support on M_0, we have that $C_6 < \infty$.

Now using Markov's inequality as in the proof of Lemma 5.9, that is, the first line of (5.34) with $p_{B',0}$ in the role of $p_{\mu'}$, gives, for any $A_2 \in \mathbb{R}$,

$$
\mathbb{P}_{\mu_1}\left(\log\frac{p_{\mu_1}(X^n)}{p_{B',0}(X^n)} < A_2\right) \leq e^{\frac{1}{2}A_2}\mathbb{E}_{\mu_1}\left[\left(\frac{p_{B',0}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right]. \tag{5.40}
$$

The expectation on the right can be further bounded, defining $\omega'' = \sqrt{\omega}/C_6$ and noting that ω'' is a probability density, as

$$
\begin{aligned}
\mathbb{E}_{\mu_1}\left[\left(\frac{p_{B',0}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right] &\leq \mathbb{E}_{\mu_1}\left[\left(\frac{\int_{\nu\in\Pi'(M'_0)}\omega(\nu)^{1/2}p_\nu(X^n)^{1/2}d\nu}{p_{\mu_1}(X^n)^{1/2}}\right)\right]\\
&= C_6\cdot\mathbb{E}_{\mu\sim\omega''}\mathbb{E}_{\mu_1}\left[\left(\frac{p_\mu(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right] \leq C_6\cdot\mathbb{E}_{\mu_1}\left[\left(\frac{p_{\mu^\circ}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right],
\end{aligned}
$$

where $\mu^\circ \in M'_0$ achieves the supremum of $E_{\mu_1}\left[\left(\frac{p_{\mu^\circ}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right]$ within M'_0. By compactness of M'_0 and continuity, this supremum is achieved. The final term can be rewritten, following the same steps as in the second and third line of (5.34), as

$$
\mathbb{E}_{\mu_1}\left[\left(\frac{p_{\mu^\circ}(X^n)}{p_{\mu_1}(X^n)}\right)^{\frac{1}{2}}\right] = e^{-\frac{n}{2}d_R(\mu_1,\mu^\circ)}. \tag{5.41}
$$

Since M'_0 and M'_1 are both CINECSI, it now follows from Proposition 5.2 that for some fixed $C_7 > 0$,

$$
d_R(\mu_1,\mu^\circ) \geq C_7\|\mu_1 - \mu^\circ\|_2^2 \geq C_7\|\mu_1 - \mu_0\|_2^2, \tag{5.42}
$$

where the latter inequality follows by the definition of $\mu_0 = \Pi_0(\mu_1)$, see the explanation below (5.11). Combining (5.40), (5.41) and (5.42), we have thus shown that for all n, all

$\mu_1 \in M_1$, all $A_2 \in \mathbb{R}$,

$$\mathbb{P}_{\mu_1}\left(\log \frac{p_{\mu_1}(X^n)}{p_{B',0}(X^n)} < A_2\right) \leq C_6 e^{\frac{1}{2}A''} e^{-\frac{n}{2}C_7 \|\mu_1 - \mu_0\|_2^2}. \tag{5.43}$$

The result now follows by combining (5.37), (5.38) and (5.43). □

5.7.4 Proof of main result, Theorem 5.5

Proof Idea The proof is based on analyzing what happens if $X_1, X_2, \ldots, X_n$ are sampled from $p_{\mu_1^{(n)}}$, where $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ are a sequence of parameters in M_1'. We consider three regimes, depending on how fast (if at all) $\mu_1^{(n)}$ converges to $\mu_0^{(n)}$ as $n \to \infty$. Here $\mu_0^{(n)} = \Pi_0(\mu_1^{(n)})$ is the projection of $\mu^{(1)}$ onto M_0, i.e. the distribution in $\mathcal{M}_0$ defined, for each n, as in (5.11), with μ_1 and μ_0 in the role of $\mu_1^{(n)}$ and $\mu_0^{(n)}$, respectively. Our regimes are defined in terms of the function f given by

$$f(n) := \frac{\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2}{\frac{\log\log n}{n}} = \frac{n \cdot \|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2}{\log\log n}, \tag{5.44}$$

which indicates how fast $d_{SQ}(\mu_1^{(n)}, \mu_0^{(n)})$ grows relative to the best possible rate $(\log\log n)/n$. We fix appropriate constants Γ_1 and Γ_2, and we distinguish, for all n with $\Gamma_2 \log n \geq \Gamma_1$, the cases:

$$f(n) \in \begin{cases} [0, \Gamma_1] & \text{Case 1} \\ [\Gamma_1, \Gamma_2 \log n] & \text{Case 2 (Theorem 5.14)} \\ [\Gamma_2 \log n, \infty] & \text{Case 3 (Theorem 5.13).} \end{cases}$$

For Case 1, the rate is easily seen to be upper bounded by $O((\log\log n)/n)$, as shown inside the proof of Theorem 5.5. In Case 3, Theorem 5.14 establishes that the probability that model $\mathcal{M}_0$ is chosen is at most of order $1/(\log n)$, which, as shown inside the proof of Theorem 5.5, again implies an upper-bound on the rate-of convergence of $O((\log\log n)/n)$. Theorem 5.13 shows that in Case 3, which includes the case that $\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2$ does not converge at all, the probability that model $\mathcal{M}_0$ is chosen is at most of order $1/n$, which, as again shown inside the proof of Theorem 5.5, again implies an upper-bound on the rate-of convergence of $O((\log\log n)/n)$.

The two theorems take into account that $\mu_1^{(n)}$ is not just a fixed function of n, but may in reality be chosen by nature in a worst-case manner, and that $f(n)$ may actually fluctuate between regions for different n. Combining these two results, we finally prove the main theorem, Theorem 5.5.

Theorem 5.13. Let M_0, M_1, M_1' and $p_{\mathrm{sw},1}(x^n)$ be as in Theorem 5.5. Then there exist positive constants $\vec{b}, \vec{c}$ such that for all $\mu_1 \in M_1'$, all n,

$$\mathbb{P}_{\mu_1}\left(\delta_{\mathrm{sw}}(X^n) = 0\right) \leq c_1 \cdot n^{m_1/2} \cdot e^{-c_2 n \|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2} + \mathtt{small}_{\vec{b}}(n), \tag{5.45}$$

where $\mu_0^{(n)} = \Pi_0(\mu_1^{(n)})$ is as in (5.11). As a consequence, with $\Gamma_2 := c_2^{-1}(1 + m_1/2)$, we have the following: for every sequence $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ with $f(n)$ as in (5.44) larger than $\Gamma_2 \log n$, we have

$$\mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) \leq \frac{c_1}{n} + \mathtt{small}_{\vec{b}}(n).$$

Proof. We can bound the probability of selecting the simple model by:

$$\begin{aligned}\mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) &= \mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)} \leq 1\right) = \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\sum_{i=0}^{\infty} \pi(2^i)\bar{p}_{2^i}(X^n)}{p_{B,0}(X^n)} \leq 1\right) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\pi(1)p_{B,1}(X^n)}{p_{B,0}(X^n)} \leq 1\right).\end{aligned}$$

Now (5.45) follows directly by applying Lemma 5.12 to the rightmost probability. For the second part, set $\Gamma_2 = c_2^{-1}(1+m_1/2)$. By assumption $f(n) > \Gamma_2 \log n$, we have $\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2 > \Gamma_2(\log n)(\log\log n)/n$. Applying (5.45) now gives the desired result. □

Theorem 5.14. Let f be as in (5.44) and M_1' be as in Theorem 5.5. For any $\gamma > 0$, there exist constants $\Gamma_1, \Gamma_3 > 0$ such that, for every sequence $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$ of elements of M_1' with for all n, $f(n) > \Gamma_1$, we have

$$\mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)} \leq \gamma\right) \leq \frac{\Gamma_3}{\log n}. \tag{5.46}$$

In particular, by taking $\gamma = 1$, we have

$$\mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) \leq \frac{\Gamma_3}{\log n}.$$

The probabilities thus converge uniformly at rate $O(1/(\log n))$ for all such sequences $\mu_1^{(1)}, \mu_1^{(2)}, \ldots$.

Proof. We specify Γ_1 later. By assumption, we have $\pi(2^i) \gtrsim (\log n)^{-\kappa}$ for $i \in \{0, \ldots, \lfloor \log_2 n \rfloor\}$. We can restrict our attention to the strategy that switches to the complex model at the penultimate switching index, due to the following inequality: for any fixed γ, there exist positive constants $\vec{C}$ such that for all large n:

$$\begin{aligned}\mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)} \leq \gamma\right) &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\sum_{i=0}^{\lfloor \log_2 n \rfloor} \pi(2^i)\bar{p}_{2^i}(X^n)}{p_{B,0}(X^n)} \leq \gamma\right) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\sum_{i=0}^{\lfloor \log_2 n \rfloor} \bar{p}_{2^i}(X^n)}{p_{B,0}(X^n)} \leq C_1(\log n)^{\kappa}\right) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\frac{\bar{p}_{2^{\lfloor \log_2 n \rfloor - 1}}(X^n)}{p_{B,0}(X^n)} \leq C_1(\log n)^{\kappa}\right) \\ &= \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{\bar{p}_{2^{\lfloor \log_2 n \rfloor - 1}}(X^n)}{p_{B,0}(X^n)} \leq \kappa \log\log n + C_2\right).\end{aligned} \tag{5.47}$$

For the remainder of this proof, we will denote the penultimate switching index by n^*, that is: $n^* = 2^{\lfloor \log_2 n \rfloor - 1}$. Now apply Lemma 5.11 twice, which gives that there exist C_3, C_4 such that, with probability at least $1 - \texttt{small}_{\vec{b}}(n)$,

$$\begin{aligned}
\log \bar{p}_{n^*}(X^n) &= \log p_{B,0}(X^{n^*}) + \log p_{B,1}(X^n | X^{n^*}) = \\
&= \log p_{B,0}(X^{n^*}) + \log p_{B,1}(X^n) - \log p_{B,1}(X^{n^*}) \\
&\geq \log p_{B,0}(X^{n^*}) + \log p_{\widehat{\mu}_1(X^n)}(X^n) - \log p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*}) + \frac{m_1}{2} \log \frac{n^*}{n} - C_3 \\
&\geq \log p_{B,0}(X^{n^*}) + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_4, \qquad (5.48)
\end{aligned}$$

where we used that $\log \frac{n^*}{n}$ is of the order of a constant, because n^* is between $\frac{n}{4}$ and $\frac{n}{2}$. From this, applying again Lemma 5.11 twice, it follows that there exists $\vec{b}$ and C_5, C_6 such that for all n, with probability at least $1 - \texttt{small}_{\vec{b}}(n)$,

$$\begin{aligned}
\log \frac{\bar{p}_{n^*}(X^n)}{p_{B,0}(X^n)} &\geq \log \frac{p_{B,0}(X^{n^*})}{p_{B,0}(X^n)} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_4 \\
&= -\log \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} - \frac{m_0}{2} \log \frac{n^*}{n} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_5 \\
&\geq -\log \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_6 \qquad (5.49)
\end{aligned}$$

where we again used that $\log \frac{n^*}{n}$ can be bounded by constants. Let $\mathcal{B}_n$ be the event that (5.49) holds. By (5.47) and (5.49), for all large n, all $\beta \geq 1$,

$$\begin{aligned}
&\mathbb{P}_{\mu_1^{(n)}}\left(\frac{p_{\mathrm{sw},1}(X^n)}{p_{B,0}(X^n)} \leq \gamma\right) \leq \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{\bar{p}_{n^*}(X^n)}{p_{B,0}(X^n)} \leq \kappa \log\log n + C_2\right) \\
\leq &\mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{\bar{p}_{n^*}(X^n)}{p_{B,0}(X^n)} \leq \kappa \log\log n + C_2, \mathcal{B}_n\right) + \mathbb{P}_{\mu_1^{(n)}}\left(\mathcal{B}_n^c\right) \\
\leq &\mathbb{P}_{\mu_1^{(n)}}\left(-\log \frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} + \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} - C_6 \leq \kappa \log\log n + C_2\right) + \texttt{small}_{\vec{b}}(n) \\
= &\mathbb{P}_{\mu_1^{(n)}}\left(\mathcal{E}_n^{(1)}\right) + \texttt{small}_{\vec{b}}(n) \leq \mathbb{P}_{\mu_1^{(n)}}\left(\mathcal{E}_n^{(\beta)}\right) + \texttt{small}_{\vec{b}}(n), \qquad (5.50)
\end{aligned}$$

where we defined

$$\mathcal{E}_n^{(\beta)} = \left\{ \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} \cdot \frac{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})}{p_{\widehat{\mu}_0(X^n)}(X^n)} \leq A_n^{(\beta)} \right\} \qquad (5.51)$$

and, for $\beta \geq 1$, we set $A_n^{(\beta)} = \beta\kappa \log\log n + C_2 + C_6$.

Below, if a sample is split up into two parts $x_1, \ldots, x_{n^*}$ and $x_{n^*+1}, \ldots, x_n$, these partial samples will be referred to as x^{n^*} and $x^{>n^*}$ respectively. We also suppress in our notation

the dependency of A_n, $\mathcal{E}_n$ and $\mathcal{D}_{j,n}$ as defined below on β; all results below hold, with the same constants, for any $\beta \geq 1$.

We will now bound the right-hand side of (5.50) further. Define the events

$$\mathcal{D}_{1,n} = \left\{ \log \frac{p_{\mu_1^{(n)}}(x^n)}{p_{\mu_1^{(n)}}(x^{n^*})} \leq \log \frac{p_{\widehat{\mu}_1(X^n)}(x^n)}{p_{\widehat{\mu}_1(X^{n^*})}(x^{n^*})} + A_n \right\}$$
$$\mathcal{D}_{0,n} = \left\{ \log \frac{p_{\mu_0^{(n)}}(x^n)}{p_{\mu_0^{(n)}}(x^{n^*})} \geq \log \frac{p_{\widehat{\mu}_0(X^n)}(x^n)}{p_{\widehat{\mu}_0(X^{n^*})}(x^{n^*})} - A_n \right\}.$$

The probability in (5.50) can be bounded, for all $\beta \geq 1$, as

$$\begin{aligned}\mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n) &= \mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, \mathcal{D}_{0,n} \cap \mathcal{D}_{1,n}) + \mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, (\mathcal{D}_{0,n} \cap \mathcal{D}_{1,n})^c) + \texttt{small}_{\vec{b}}(n) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, \mathcal{D}_{0,n}, \mathcal{D}_{1,n}) + \mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{1,n}^c) + \mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{0,n}^c) + \texttt{small}_{\vec{b}}(n). \end{aligned} \tag{5.52}$$

We first consider the first probability in (5.52): there are constants $\vec{C}$ such that, for all large n,

$$\begin{aligned}&\mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n, \mathcal{D}_{0,n}, \mathcal{D}_{1,n}) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\mu_1^{(n)}}(X^n)}{p_{\mu_1^{(n^*)}}(X^{n^*})} - A_n + \log \frac{p_{\mu_0^{(n)}}(X^n)}{p_{\mu_0^{(n)}}(X^{n*})} - A_n \leq A_n \right) \\ &= \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\mu_1^{(n)}}(X^{>n^*})}{p_{\mu_0^{(n)}}(X^{>n^*})} \leq 3A_n \right) \\ &\leq e^{\frac{3}{2}A_n} e^{-\frac{n}{4} d_R(\mu_1^{(n)}, \mu_0^{(n)})} \leq e^{(3/2)\beta\kappa \log\log n + C_7} e^{-C_8 n \|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2} = e^{C_7} (\log n)^{(3/2)\beta\kappa - \Gamma_1 \cdot C_8}, \end{aligned} \tag{5.53}$$

where Γ_1 is as in the statement of the theorem, the second inequality follows by Lemma 5.9 and noting $n^* < \frac{n}{2}$, we used Proposition 5.2.

We now consider the second probability in (5.52). Using $p_{\widehat{\mu}_1(X^n)}(x^n) \geq p_{\mu_1^{(n)}}(x^n)$ we have the following, where we define the event $\mathcal{F}_n = \{\widehat{\mu}_1(X^{n^*}) \in M_1'\}$ with M_1' the CINECSI subset of M_1 mentioned in the theorem statement: there is $C_9, C_{10} > 0$ such that for al large n,

$$\begin{aligned}\mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{1,n}^c) &= \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\mu_1^{(n)}}(X^n)}{p_{\mu_1^{(n)}}(X^{n^*})} > \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} + A_n \right) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\mu_1^{(n)}}(X^{n^*})} > \log \frac{p_{\widehat{\mu}_1(X^n)}(X^n)}{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})} + A_n \right) \\ &\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log \frac{p_{\widehat{\mu}_1(X^{n^*})}(X^{n^*})}{p_{\mu_1^{(n)}}(X^{n^*})} > A_n, \mathcal{F}_n \right) + \mathbb{P}_{\mu_1^{(n)}}\left(\mathcal{F}_n^c\right) \end{aligned}$$

$$\leq \mathbb{P}_{\mu_1^{(n)}}\left(D(\widehat{\mu}_1(X^{n^*})\|\mu_1^{(n)}) > A_n, \mathcal{F}_n\right) + \texttt{small}_{\vec{b}}(n)$$

$$\leq \mathbb{P}_{\mu_1^{(n)}}\left(\|\widehat{\mu}_1(X^{n^*}) - \mu_1^{(n)}\|_2^2 > C_9 A_n, \mathcal{F}_n\right) + \texttt{small}_{\vec{b}}(n)$$

$$\leq \mathbb{P}_{\mu_1^{(n)}}\left(\|\widehat{\mu}_1(X^{n^*}) - \mu_1^{(n)}\|_\infty > \sqrt{C_9 A_n/m_1}\right) + \texttt{small}_{\vec{b}}(n) \tag{5.54}$$

$$\leq e^{-C_{10}A_n} = e^{-C_{10}(C_2-C_6)}\frac{1}{(\log n)^{C_{10}\beta\kappa}}, \tag{5.55}$$

where we used the KL robustness property (5.25), Proposition 5.2 and Lemma 5.8.

The third probability in (5.52) is considered in a similar way. Using $p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*}) \geq p_{\mu_0^{(n)}}(X^{n^*})$ we have $C_{11}, C_{12} > 0$ such that:

$$\begin{aligned}
\mathbb{P}_{\mu_1^{(n)}}(\mathcal{D}_{0,n}^c) &= \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\mu_0^{(n)}}(X^n)}{p_{\mu_0^{(n)}}(X^{n^*})} < \log\frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} - \frac{1}{3}A_n\right) \\
&\leq \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\mu_0^{(n)}}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(x^{n^*})} < \log\frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\widehat{\mu}_0(X^{n^*})}(X^{n^*})} - \frac{1}{3}A_n\right) \\
&= \mathbb{P}_{\mu_1^{(n)}}\left(\log\frac{p_{\widehat{\mu}_0(X^n)}(X^n)}{p_{\mu_0^{(n)}}(X^n)} > \frac{1}{3}A_n\right) \\
&\leq C_{11}\frac{1}{(\log n)^{C_{12}\beta\kappa}}
\end{aligned} \tag{5.56}$$

where we omitted the last few steps which are exactly as in (5.54).

We now finish the proof by combining (5.52), (5.53), (5.54) and (5.56), which gives that, if we choose $\beta \geq \max\{1/(\kappa C_{10}), 1/(\kappa C_{12})\}$ and, for this choice of β, we choose Γ_1 as in (5.53) as $\Gamma_1 \geq (1 + (3/2)\beta\kappa)/C_8$, then we have $\mathbb{P}_{\mu_1^{(n)}}(\mathcal{E}_n) \leq \Gamma_4/(\log n)$ for some constant Γ_4 independent of n; the result now follows from (5.50).

□

Proof of Theorem 5.5

Proof. We show the result in two stages. In Stage 1 we provide a tight upper bound on the risk, based on an extension of the decomposition of the risk (5.12) to general families and estimators $\breve{\mu}_0$ and $\breve{\mu}_1$ that are sufficiently efficient, i.e. that satisfy (5.13), and to losses $d_{\text{gen}}(\cdot\|\cdot)$ equal to squared error loss, standardized squared error loss and KL divergence (it is not sufficient to refer to Proposition 5.2 and prove the result only for squared error loss, because the equivalence result of Proposition 5.2 only holds on CINECSI sets and our estimators may take values outside of these; we do not need to consider Rényi and squared Hellinger divergences though, because these are uniformly upper bounded by KL divergence even for μ outside any CINECSI set). In Stage 2 we show how the bound implies the result.

Stage 1: Decomposition of upper bound on the risk Let A_n be the event that $\mathcal{M}_1$ is selected, as in Section 5.3.3. We will now show that, under the assumptions of Theorem 5.5,

we have for the constant C appearing in (5.13), for all $\mu_1 \in M_1'$,

$$R(\mu_1, \delta, n) \leq \frac{3C}{n} + 2\mathbb{P}(A_n^c) d_{\mathrm{gen}}(\mu_1 \| \mu_0), \tag{5.57}$$

where the left inequality holds for all divergence measures mentioned in the theorem, and the right inequality holds for $d_{\mathrm{gen}}(\cdot\|\cdot)$ set to any of the squared error, the standardized squared error or the KL divergence.

To prove (5.57), we use that for the three divergences of interest, for any $\mu_1 \in M_1, \mu \in M_0$, with $\mu_0 \in M_0$ as in (5.11), we have

$$d_{\mathrm{gen}}(\mu_1 \| \mu) \leq 2(d_{\mathrm{gen}}(\mu_1 \| \mu_0) + d_{\mathrm{gen}}(\mu_0 \| \mu)), \tag{5.58}$$

For $d_{\mathrm{gen}}(\cdot\|\cdot)$ the KL divergence, this follows because

$$\begin{aligned} D(\mu_1 \| \mu) &= \mathbb{E}_{\mu_1}\left[-\log \frac{p_\mu(X)}{p_{\mu_1}(X)}\right] = \mathbb{E}_{\mu_1}\left[-\log \frac{p_\mu(X)}{p_{\mu_0}(X)}\right] + \mathbb{E}_{\mu_1}\left[-\log \frac{p_{\mu_0}(X)}{p_{\mu_1}(X)}\right] \\ &= \mathbb{E}_{\mu_0}\left[-\log \frac{p_\mu(X)}{p_{\mu_0}(X)}\right] + \mathbb{E}_{\mu_1}\left[-\log \frac{p_{\mu_0}(X)}{p_{\mu_1}(X)}\right], \end{aligned} \tag{5.59}$$

where the last line follows by the robustness property of exponential families (5.24), since μ and μ_0 are both in M_0.

For $d_{\mathrm{gen}}(\cdot\|\cdot)$ the squared and standardized squared error case we show (5.58) as follows: Fix a matrix-valued function $J : M_1 \to \mathbb{R}^{m_1^2}$ that maps each $\mu \in M_1$ to a positive definite matrix J_μ. We can write

$$d_{\mathrm{gen}}(\mu \| \mu') = (\mu - \mu')^T J_\mu (\mu - \mu'). \tag{5.60}$$

where J_μ is the identity matrix for the squared error case, and J_μ is the Fisher information matrix for the standardized squared error case. (5.58) follows since we can write, for any function J_μ of the above type including these two:

$$\begin{aligned} (\mu_1 - \mu)^T J_{\mu_1}(\mu_1 - \mu) &= (\mu_1 - \mu_0 + \mu_0 - \mu)^T J_{\mu_1}(\mu_1 - \mu_0 + \mu_0 - \mu) \\ &= (\mu_1 - \mu_0)^T J_{\mu_1}(\mu_1 - \mu_0) + (\mu_0 - \mu)^T J_{\mu_1}(\mu_0 - \mu) + 2(\mu_1 - \mu_0) J_{\mu_1}(\mu_0 - \mu) \\ &\leq 2\left((\mu_1 - \mu_0)^T J_{\mu_1}(\mu_1 - \mu_0) + (\mu_0 - \mu)^T J_{\mu_1}(\mu_0 - \mu)\right), \end{aligned}$$

where the last line follows because for general positive definite $m \times m$ matrices J and m-component column vectors a and b, $(b-a)^T J(b-a) \geq 0$ so that $b^T J(b-a) \geq a^T J(b-a)$ and, after rearranging, $b^T J b + a^T J a \geq 2a^T J b$.

We have thus shown (5.58). It now follows that

$$\begin{aligned} R(\mu_1, \delta, n) &= \mathbb{E}_{\mu_1}\left[\mathbf{1}_{A_n} d_{\mathrm{gen}}(\mu_1 \| \breve{\mu}_1(X^n)) + \mathbf{1}_{A_n^c} d_{\mathrm{gen}}(\mu_1 \| \breve{\mu}_0(X^n))\right] \\ &\leq \mathbb{E}_{\mu_1}\left[d_{\mathrm{gen}}(\mu_1 \| \breve{\mu}_1(X^n)) + 2 \cdot \mathbf{1}_{A_n^c}\left(d_{\mathrm{gen}}(\mu_0 \| \breve{\mu}_0(X^n)) + d_{\mathrm{gen}}(\mu_1 \| \mu_0)\right)\right] \\ &\leq \frac{3C}{n} + 2\mathbb{P}(A_n^c) d_{\mathrm{gen}}(\mu_1 \| \mu_0), \end{aligned} \tag{5.61}$$

where we used (5.58) and our condition (5.13) on $\breve{\mu}_0$ and $\breve{\mu}_1$. We have thus shown (5.57).

Stage 2 We proceed to prove our risk upper bound for the squared error loss, standardized squared error loss and KL divergence, for which the right inequality in (5.57) holds; the result then follows for squared Hellinger and Rényi divergence because these are upper bounded by KL divergence. From (5.57) we see that it is sufficient to show that for all n larger than some n_0,

$$\sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(A_n^c) d_{\mathrm{gen}}(\mu_1 \| \mu_0)\} = O\left(\frac{\log\log n}{n}\right), \tag{5.62}$$

for our three choices of $d_{\mathrm{gen}}(\cdot\|\cdot)$. We first note that, since M_1' is CINECSI, $\sup_{\mu_1 \in M_1'} d_{\mathrm{gen}}(\mu_1\|\mu_0)$ is bounded by some constant C_1. It thus follows by Proposition 5.10 that there exists some CINECSI subset M_1'' of M_1 such that, with $B_n^c \subset A_n^c$ defined as $B_n^c = \{x^n : \delta(x^n) = 0; \widehat{\mu}_1(X^n) \in M_1''\}$, we have

$$\begin{aligned}
\sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(A_n^c) d_{\mathrm{gen}}(\mu_1\|\mu_0)\} &= \sup_{\mu_1 \in M_1'} \{(\mathbb{P}_{\mu_1}(B_n^c) + \mathbb{P}_{\mu_1}(A_n^c \setminus B_n^c)) d_{\mathrm{gen}}(\mu_1\|\mu_0)\} \\
&= \sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(B_n^c) d_{\mathrm{gen}}(\mu_1\|\mu_0)\} + C_1 \cdot \mathbb{P}_{\mu_1}(\widehat{\mu}^{(1)} \notin M_1'') \\
&= \sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(B_n^c) d_{\mathrm{gen}}(\mu_1\|\mu_0)\} + \mathtt{small}_{\vec{b}}(n),
\end{aligned}$$

so that it is sufficient if we can show (5.62) with B_n^c instead of A_n^c. But on the set B_n^c, all three divergence measures considered are within constant factors of each other, so that it is sufficient if we can show that there is a constant C_2 such that for all n larger than some n_0,

$$\sup_{\mu_1 \in M_1'} \{\mathbb{P}_{\mu_1}(B_n^c) \cdot \|\mu_1 - \mu_0\|_2^2\} \le C_2 \cdot \frac{\log\log n}{n}. \tag{5.63}$$

Now, fix some $\mu_1 \equiv \mu_1^{(n)}$ and consider $f(n)$ as in (5.44). By Theorem 5.13, $\mathbb{P}_{\mu_1}(B_n^c) \le C_3/n$ for some constant C_3 that can be chosen uniformly for all $\mu_1 \in M_1'$ whenever $f(n) > \Gamma_2 \log n$ with Γ_2 as in that theorem. Using also that $\|\mu_1 - \mu_0\|_2^2$ is bounded by C_1 as above, it follows that (5.63) holds whenever $f(n) > \Gamma_2 \log n$ and $(C_1C_3)/n \le C_2(\log\log n)/n$, i.e. whenever $f(n) > \Gamma_2 \log n$ and $C_2 \ge C_1C_3/(\log\log n)$.

Second, suppose that $\Gamma_1 < f(n) \le \Gamma_2 \log n$ with Γ_1 as in Theorem 5.14. Then by that theorem, uniformly for all $\mu_1^{(n)}$ with such $f(n)$, we have, with Γ_3 as in that theorem,

$$\begin{aligned}
&\|\mu_1^{(n)} - \mu_0^{(n)}\|_2^2 \cdot \mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) = f(n) \cdot \frac{\log\log n}{n} \cdot \mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) \le \\
&\Gamma_2 \cdot (\log n) \cdot \frac{\log\log n}{n} \cdot \mathbb{P}_{\mu_1^{(n)}}(\delta_{\mathrm{sw}}(X^n) = 0) \le \Gamma_2\Gamma_3 \cdot \frac{\log\log n}{n},
\end{aligned}$$

where $\mu_0^{(n)} = \Pi_0(\mu_1^{(n)})$ is defined as in (5.11), so that (5.63) holds again whenever $C_2 \ge \Gamma_2\Gamma_3$.

Finally, suppose that $f(n) \le \Gamma_1$ with Γ_1 as in Theorem 5.14. Then (5.63) holds whenever $C_2 \ge \Gamma_1$. Combining the three cases we find that (5.63) holds whenever $C_3 \ge \max\{\Gamma_1, \Gamma_2\Gamma_3, C_1C_3/(\log\log n)\}$; the result is proved.

□

5.7.5 Switching as in Van Erven et al. (2012)

The basic building block of the switch distribution and criterion as formulated by Van Erven et al. (2012) is a countable set of *sequential prediction strategies* (also known as 'prequential forecasting systems' (Dawid, 1984)) $\{p_k \mid k \in \mathcal{K}\}$, where $\mathcal{K}$ is a finite or countable set indexing the basic models under consideration. Thus, each model is associated with a corresponding prediction strategy, where a prediction strategy p is a function from $\bigcup_{i\geq 0} \mathcal{X}^i$ to the set of densities on $\mathcal{X}$, where $p(\cdot \mid x^{n-1})$ denotes the density on $\mathcal{X}$ that x^{n-1} maps to, and $p(x_n \mid x^{n-1})$ is to be interpreted as the probabilistic prediction that strategy p makes for outcome X_n upon observation of the first $n-1$ outcomes, $X^{n-1} = x^{n-1}$. For example, for a parametric model $\{p_\theta \mid \theta \in \Theta\}$ one can base p_k on a Bayesian marginal likelihood, $p_B(x^n) := \int_\Theta \omega(\theta) p_\theta(x^n) d\theta$, where ω is a prior density on Θ. The corresponding prediction strategy could then be defined by setting $p_k(x_n \mid x^{n-1}) := p_B(x^n)/p_B(x^{n-1})$, the standard Bayesian predictive distribution. In this paper, the basic strategies p_k were always Bayesian predictive distributions, but, in the spirit of Dawid (1984), one may consider other choices as well.

After constructing the set of basic prediction strategies, a new family of prediction strategies that switch between the strategies in the set $\{p_k \mid k \in \mathcal{K}\}$ is defined. Formally, let $\mathbb{S}$ be the set

$$\mathbb{S} = \{((t_1,k_1),\ldots,(t_m,k_m)) \in (\mathbb{N}\times\mathcal{K})^m \mid m \in \mathbb{N}, 1 = t_1 < t_2 < \ldots < t_m\}. \tag{5.64}$$

Each $s \in \mathbb{S}$ specifies the times $t_1,\ldots,t_m$ at which a switch is made between the prediction strategies from the original set, identified by the indices $k_1,\ldots,k_m$. The new family $Q = \{q_s \mid s \in \mathbb{S}\}$ is then defined by setting, for all $n, x^n \in \mathcal{X}^n$:

$$q_s(x_n \mid x^{n-1}) = p_{k_j}(x_n \mid x^{n-1}), \quad t_j \leq n < t_{j+1}, \tag{5.65}$$

with $t_{m+1} = \infty$ by convention. We now define $q_s(x^n) = \prod_{i=1}^n q_s(x_i \mid x^{i-1})$; one easily verifies that this defines a joint probability density on $\mathcal{X}^n$.

We now place a prior mass function π' on $\mathbb{S}$ and define, for each n, the *switch distribution* in terms of its joint density for $\mathcal{X}^n$ and $\mathbb{S}$:

$$p_{\mathrm{sw}}(x^n,s) = q_s(x^n)\pi'(s), \quad p_{\mathrm{sw}}(x^n) = \sum_{s\in\mathbb{S}} p_{\mathrm{sw}}(x^n,s) = \sum_{s\in\mathbb{S}} q_s(x^n)\pi'(s).$$

If the p_k are defined as Bayesian predictive distributions as above, then, as explained by Van Erven et al. (2012), the density $p_{\mathrm{sw}}(x^n)$ can be interpreted as a Bayesian marginal density of x^n under the prior π' on meta-models (model sequences) in $\mathbb{S}$.

The switch distribution can be used to define a model selection criterion δ'_{sw} by selecting the model with highest posterior probability under the switch distribution. This is done by defining the random variable $K_{n+1}(s)$ on $\mathbb{S}$ to be the index of the prediction strategy that is used by q_s to predict the $(n+1)$th outcome. The model selection criterion is then:

$$\delta'_{\mathrm{sw}}(x^n) = \arg\max_k p_{\mathrm{sw}}(K_{n+1} = k \mid x^n) = \arg\max_k \frac{\sum_{s:K_{n+1}(s)=k} p_{\mathrm{sw}}(x^n,s)}{p_{\mathrm{sw}}(x^n)}$$

$$= \arg\max_k \frac{\sum_{s:K_{n+1}(s)=k} q_s(x^n)\pi'(s)}{\sum_{s\in\mathbb{S}} q_s(x^n)\pi'(s)}, \tag{5.66}$$

with ties resolved in any way desired.

In our nested two-model case, one might use, for example, a prior π' with support on

$$\mathbb{S}' = \{(1,0),(1,1),(\ (1,0),(2,1)\ ),(\ (1,0),(4,1)\ ),(\ (1,0),(8,1)\ ),(\ (1,0),(16,1),\dots)\}.$$

Such a prior expresses that at time 1, for the first prediction, one can either switch to (i.e., start with), model 0, and keep predicting according to its Bayes predictive distribution — this strategy gets weight $\pi((1,0))$. Or one can start with model 1, and keep predicting according to its Bayes predictive distribution — this strategy gets weight $\pi((1,1))$. Or one can start with model 0 and switch to model 1 after 2^i observations and then stick with 1 forever — this strategy gets weight $\pi((\ (1,0),(2^i,1)\ ))$. If we now start with a prior π on $\{1,2,\dots\}$ as in the main text and define $\pi'((1,0)) = 1/2$, $\pi'((1,1)) = (1/2)\cdot\pi(1)$, and for $i \geq 1$, $\pi'((1,0),(2^i,1)) = (1/2)\cdot\pi(2^i)$, then $\sum_{s\in\mathbb{S}'}\pi'(s) = 1$, so π' is a probability mass function. A simple calculation gives that (5.66) based on switch prior π' now chooses model 1 if

$$\sum_{1\leq t<n} \bar{p}_t(x^n)\pi(t) > (1+g(n))\cdot p_{B,0}(x^n), \tag{5.67}$$

where $g(n) = \sum_{t\geq n}\pi(t)$; note that $g(n)$ is decreasing and converges to 0 with increasing n. (5.67) is thus an instance of the switch criterion of Van Erven et al. (2012). Comparing this to (5.3), the criterion used in this paper, after rearranging we see that it chooses model 1 if

$$\sum_{1\leq t<n} \bar{p}_t(x^n)\pi(t) > (1-g(n))\cdot p_{B,0}(x^n),$$

which is more likely by constant factor to select model $\mathcal{M}_0$, the factor however tending to 1 with increasing n. It is completely straightforward to check that Theorem 5.5 and all other results in this paper still hold if δ_{sw} with prior π as in the main text is replaced by δ'_{sw} with corresponding prior π' as defined here; thus our results carry over to the original definitions of Van Erven et al. (2012). Similarly, the proof for the strong consistency of δ'_{sw} given by Van Erven et al. (2012) carries through for δ_{sw}, needing only trivial modifications.

6

Bilateral patients in arthroplasty registry data

6.1 Introduction

Worldwide more than 3 million total hip and knee arthroplasties are performed annually, and this number is predicted to increase substantially within the next decades (Pabinger and Geissler, 2014; Pabinger et al., 2015). Data on total joint arthroplasties (TJAs) are collected in a growing number of arthroplasty registries around the world, and the resulting data has proven to be valuable in improving the outcome of TJA (Graves, 2010).

This chapter is based on analyses of total hip arthroplasty (THA) data from the LROI (Landelijke Registratie Orthopedische Implantaten / Dutch Arthroplasty Register), which has been recording patient and implant characteristics of all hip and knee replacements in The Netherlands since its establishment in 2007. A large number of THAs is performed in The Netherlands each year; the LROI registered about 28.000 primary THRs in 2014. Osteoarthritis is the most common reason for THA: 87% of THAs were performed after a diagnosis of osteoarthritis (LROI, 2014).

Benefits of THA include improved mobility, increased hip joint functionality, and pain relief (Wilcock, 1978). A hip implant does not last forever however, and a patient may

This chapter contains material from two papers. The first has been submitted as: S.L. van der Pas, R.G.H.H. Nelissen and M. Fiocco. Staged bilateral total joint arthroplasty patients in registries. Immortal time bias and methodological options. The second is in preparation, with R.G.H.H. Nelissen, B.W. Schreurs and M. Fiocco, and titled 'Risk factors for early revision after unilateral and staged bilateral total hip replacement in the Dutch Arthroplasty Register'.

need to undergo revision surgery, which we define as any change to the implant. Revision places not only a burden on healthcare costs, but on the patient as well, and is associated with higher risk of adverse outcomes than the primary surgery (Mahomed et al., 2003; Ong et al., 2006). Incidence of revision has been linked to many demographic, clinical, surgical and health care provider related factors, including gender, age, race, body weight, American Society of Anesthesiologists (ASA) score, underlying diagnosis, type of fixation and hospital volume (Prokopetz et al., 2012). We investigate risk factors for revision within the first 8 years of follow-up.

Three methodological issues need to be taken into consideration during the statistical analysis. The first and second are due to the presence of (staged) bilateral patients in the data. With "bilateral patients", we refer to patients with two THAs, and we refer to patients with a THA on one side as "unilateral patients". The first issue is that each bilateral THA patient contributes two dependent observations, violating the independence assumptions underlying most methods. Secondly, the time that usually passes between two successive THAs renders a patient's bilaterality status time-dependent. The number of patients with bilateral THAs is not negligible; in The Netherlands 20% of total hip arthroplasty surgeries in 2014 concerned the placement of a second prosthesis, in Sweden 20.5% of patients became staged bilateral between 1992-2014, and in Norway, 23.6% of patients became bilateral within 10 years (Lie et al., 2004; LROI, 2014; SHAR, 2014).

The third issue is that a patient may die before experiencing revision of the implant. If this competing risk of death is not appropriately accounted for, the risk of revision surgery will be overestimated (Keurentjes et al., 2012; Ranstam et al., 2011).

Although this chapter is written in the context of total hip replacement, the considerations and results are relevant to registry data of any body part of which a human has at least two, such as knees, ankles, shoulders, eyes, fingers and teeth.

The structure of this chapter is as follows. Methods for handling the competing risk of death are briefly reviewed in Section 6.2. The complications stemming from the bilateral patients are discussed in Section 6.3. The data structure is then introduced in Section 6.4. This Chapter concludes with preliminary results on the LROI data in Section 6.5.

6.2 Competing risk of death

THA is most commonly done in elderly patients; the average age of the patients in the hip replacement data set is 69 years. A patient may die before experiencing revision. Indeed, out of the 161,434 hips in the data set 3,897 hips were revised, while it was not possible to observe revision for 7,179 hips due to death of the patient. Thus, death should be considered a competing risk. Estimating the probability of revision by Kaplan-Meier would be inappropriate, as it is designed for a single outcome (in this case, revision), which is possibly not observed due to censoring. Deaths are treated as censored observations and not as events. However, considering deceased patients as censored observations violates the independence of the censoring distribution assumption underlying Kaplan-Meier (Putter et al., 2007). By the independent censoring assumption, a dead patient would have the same hazard of revision as a patient who is still alive and has not experienced revision yet. Since Kaplan-Meier treats dead patients as if they could still experience revision, the probability of revision is overestimated.

In a competing risk setting, the functions of interest are the cumulative incidence functions. The cumulative incidence of cause k at time t is the probability that failure due to cause k has occurred by time t. There are methods available to estimate the cumulative incidence of any event in the competing risks setting. We consider three of them, and first introduce some notation.

We assume right-censored data. We have n observations, and each observation i has failure time T_i and censoring time C_i associated to it. Define $X_i = \min\{T_i, C_i\}$, and $\Delta_i = \mathbf{1}\{T_i \leq C_i\}$ and let $\varepsilon_i \in \{1,\ldots,K\}$ be the causes of failure, for $i = 1,\ldots,n$. Let Z_i be a $p \times 1$ bounded and time-independent covariate vector. We assume that $(X_i, \Delta_i, \Delta_i\varepsilon_i, Z_i)$ are independent and identically distributed for $i = 1,\ldots,n$. Denote the observed, distinct event times by $t_1 < t_2 < \ldots < t_m$.

With this notation, the cumulative incidence of cause k is given by:

$$F_k(t) = Pr(T \leq t, \varepsilon = k), \quad k = 1,\ldots,K.$$

A cumulative incidence function is determined by the cause-specific hazards $\lambda_k(t)$, $k = 1,\ldots,K$. The cause-specific hazard is the hazard of failing from cause $k \in \{1,\ldots,K\}$, which is in competition with the other failure causes. It is defined as

$$\lambda_k(t) = \lim_{\Delta t \downarrow 0} \frac{Pr(t \leq T \leq t + \Delta t, \varepsilon = k \mid T \geq t)}{\Delta t}.$$

The cumulative incidence can be expressed in terms of the cause-specific hazards as follows:

$$F_k(t) = \int_0^t S(u) d\Lambda_k(u), \quad k = 1,\ldots,K, \tag{6.1}$$

where $S(t) = \exp(-\sum_{k=1}^K \Lambda_k(t))$ is the overall survival function, and $\Lambda_k(t) = \int_0^t \lambda_k(u)du$ is the cumulative cause-specific hazard. In the following sections, we briefly review three methods for estimating the cumulative incidence: the Aalen-Johansen estimator, cause-specific Cox regression and Fine-Gray regression.

Aalen-Johansen estimator

The unadjusted cumulative incidence can be estimated by the Aalen-Johansen estimator (Aalen and Johansen, 1978), which was defined for more general multi-state models, but in this case reduces to (6.1) with the left-continuous Kaplan-Meier estimate for the survival function, and the Nelson-Aalen estimators for the cumulative cause-specific hazards. Denote the number of failures due to cause k at time t_i by $d_k(t_i) = \sum_{i=1}^n \mathbf{1}\{X_i = t_i, \varepsilon_i = k\}$ and the number still at risk just before time t_i by $n(t_i) = \sum_{i=1}^n \mathbf{1}\{X_i \geq t_i\}$.

The Nelson-Aalen estimators and Kaplan-Meier estimator are given by

$$\widehat{\Lambda}_k(t) = \sum_{t_i \leq t} \frac{d_k(t_i)}{n(t_i)}, \quad \widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{\sum_{k=1}^K d_k(t_i)}{n(t_i)}\right),$$

and $F_k(t)$ is estimated by $\widehat{F_k}(t) = \sum_{t_i \leq t} S(t_{i-1}) d_k(t_i)/n(t_i)$.

Cause-specific Cox regression

The Cox proportional hazards model allows a natural extension to the competing risks settings, where the cause-specific hazard for individual i and cause k is modelled as (Holt, 1978):

$$\lambda_k(t; z_i) = \lambda_{0,k}(t) e^{\beta_k^T z_i}. \tag{6.2}$$

Here, $\lambda_{0,k}(t)$ is a cause-specific baseline hazard. All cause-specific hazards are estimated separately and then combined to assess the association of the covariates to the cumulative incidence of the cause of interest. Each cause-specific hazard $\lambda_k(t; z_0)$ is estimated by censoring all individuals who failed due to a cause other than k. At each time at which an individual experiences failure due to cause k, the covariate values of this individual are compared with the covariates of all other individuals who are still event-free and in follow-up. Following Cheng et al. (1998), the cumulative incidence is estimated by plugging in the maximum partial likelihood estimate $\widehat{\beta}_k$ for β_k and the Breslow estimate $\widehat{\Lambda}_{0,k}(t)$ for the cumulative hazard:

$$\widehat{F}_k(t; z_0) = \int_0^t \widehat{S}(u; z_0) d\widehat{\Lambda}_k(u; z_0),$$

where $\widehat{S}(u; z_0) = \exp(-\sum_{k=1}^K \widehat{\Lambda}_k(u; z_0))$ and $\widehat{\Lambda}_k(u; z_0) = \widehat{\Lambda}_{0,k}(u) \exp(\widehat{\beta}_k^T z_0))$.

Fine-Gray regression

Fine-Gray regression (Fine and Gray, 1999) is a Cox model like (6.2), but for the subdistribution hazard $h_k(t; z_0)$ instead of the cause-specific hazard. The subdistribution hazard is the instantaneous risk of failing from cause k given that the individual has not failed from cause k:

$$h_k(t; Z_i) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t, \varepsilon = k \mid T \geq t \cup (T \leq t \cap \varepsilon \neq k), Z_i).$$

Fine-Gray regression is designed to model only one subdistribution hazard at the same time (Beyersmann et al., 2012, Section 5.3.4). The model is given by:

$$h_k(t; z_i) = h_{0,k}(t) e^{\beta_k^T z_i}, \tag{6.3}$$

where $h_{0,k}(t)$ is a subdistribution baseline hazard and k refers to the single cause of failure under consideration. An appealing property of the subdistribution hazard is that it satisfies

$$F_k(t; z_0) = 1 - e^{-\int_0^t h_k(u; z_0) du}. \tag{6.4}$$

The model (6.3) for the subdistribution hazard thus allows direct assessment of the relationship between a covariate and the cumulative incidence of the cause of interest. The risk set corresponding to the subdistribution hazard is counterintuitive however, as it contains those individuals who have already failed from a cause different than k, and are thus not able to fail from cause k anymore (Fine and Gray, 1999). The coefficients β in (6.3) are estimated by a weighted partial likelihood approach and the cumulative subdistribution baseline hazard is estimated by a Breslow-type estimator. A test for equality of cause-specific cumulative incidence functions is available (Gray, 1988).

Choice of method and model checking

Usually, only either cause-specific Cox regression or Fine-Gray regression is selected for adjusted analyses. This seems natural, as they estimate different quantities, although both can be used to estimate cumulative incidences. There is a relationship between the subdistribution hazard $h_k(t; z_0)$ and cause-specific hazard $\lambda_k(t; z_0)$, which follows from combining (6.1), (6.4) and differentiating with respect to t (Beyersmann and Scheike, 2013; Beyersmann and Schumacher, 2007):

$$h_k(t; z_0) = \frac{S(t; z_0)}{1 - F_k(t; z_0)} \lambda_k(t; z_0).$$

If the proportionality assumption holds for one of the hazards, the other model will thus typically be misspecified (Beyersmann and Schumacher, 2007; Latouche et al., 2013, 2007). Grambauer et al. (2010) found that the subdistribution hazards and cause-specific hazards for cause 1 are numerically quite close if a covariate has no effect on the remaining cause-specific hazards, or when there is heavy censoring.

Cause-specific Cox regression provides insight into the relationship of covariates on the hazard of, in this case, revision or death. Fine-Gray regression yields in a sense a summary, indicating the association between a covariate and the cumulative incidence of revision. Grambauer et al. (2010) and Latouche et al. (2013) recommend presenting the results from both the cause-specific Cox model and Fine-Gray regression side by side, for all causes. In any case, it is prudent to report results on model fit. There are several options available. Three aspects of the models are evaluated (Lin et al., 1993):

1. The proportional hazards assumption;
2. The functional forms of covariates in the exponent of the model;
3. The link function.

An overview of diagnostic tests for the Cox model is given in Chapter 11 of Klein and Moeschberger (2003), while Li et al. (2015) discuss a number of tests for each of the three aspects listed above for Fine-Gray regression. In addition, Andersen and Pohar Perme (2010) review methods for assessing goodness-of-fit using pseudo-values, which can be applied to the Cox model and the Fine-Gray model. Omnibus tests for all three aspects are available for the Cox model (Lin et al., 1993; McKeague et al., 2001) and the Fine-Gray model (Li et al., 2015), and an R package for the latter is under development (Li et al., 2015).

6.3 Dependence between hips and the time-dependent bilateral status

Approximately 20% of THAs undertaken in The Netherlands concern the placement of a second hip implant (LROI, 2014). Thus, the LROI data contains a sizable proportion of bilateral patients. Both hips can be placed simultaneously, but more commonly, the interoperative time is several months or years. In the latter case, the patients are referred to as "staged bilateral patients". Their presence poses a problem to the statistical analysis of

arthroplasty data, as has been recognized in the orthopaedic literature (Bryant et al., 2006; Lie et al., 2004; Ranstam et al., 2011). The focus of those papers has been on the dependence of the two observations contributed by a bilateral patient. There is little recognition for a second problem however, namely that a patient's bilaterality status is time-dependent.

We review some methods for handling the dependence between two hip implants within a patient in Section 6.3.1, and discuss methods which incorporate the time-dependent status in Section 6.3.2. We give some remarks on practical relevance in Section 6.3.3.

6.3.1 Methods for dependent observations

Most of the methods proposed in the orthopaedic literature only account for the dependence between two hips within one patient, and do not consider the time between two successive THAs. The interoperative time is not relevant for patients undergoing simultaneous bilateral hip replacement ("same-day bilateral patients"), and those are the patients we will have in mind in this section. We now review the methods that have been proposed in the orthopaedic literature. These methods are intended to be used in combination with the competing risks methods discussed in Section 6.2.

Subgroup analysis

One recommendation by Bryant et al. (2006) is to analyse patients with bilateral THA as a separate subgroup, which is done occasionaly in practice, or the bilateral patients are excluded altogether (Buchholz et al., 1985; Gillam et al., 2010; SHAR, 2014; Visuri et al., 2002). The unilateral observations will all be independent, but the bilateral patients' observations are still dependent, so the dependence issue is not completely resolved by subgroup analysis. In addition, when subgroup analysis is done with staged bilateral patients, the analysis is at risk of being affected by immortal time bias, as will be explained in Section 6.3.2.

Excluding the second joint

Bryant et al. (2006) suggested excluding the second joint, and this option is used in practice (Maurer et al., 2001; Morris, 1993; NJR, 2015). Only using each patient's first THA ensures independence of the observations used in the analysis. A disadvantage is that not all data is used, although this may not be a serious problem in arthroplasty registry studies where the amount of data can run into the hundred thousands. At first glance, a second drawback may be that the conclusions only hold for a patient's first THA and not the second, but this may actually be sensible given that the outcomes for the second implant may be different compared to the first implant.

Selecting a random joint

A third suggestion by Bryant et al. (2006) is to select a random hip for each bilateral patient, and this was previously implemented by Visuri et al. (2002). The analysis is carried out using all unilateral observations, and one randomly selected observation from each bilateral patient. In this way, alle observations in the sample are independent. However,

this raises other issues. The first problem is that the sensitivity of the results to the particular sample should be assessed. The second is that it is unclear what is being estimated. If the outcomes of a patient's first and second THA are different, the interpretation of the estimate resulting from this procedure is difficult.

Resampling techniques

Closely related to the selection of a random joint per patient is the idea of within-cluster resampling. Each patient is viewed as a cluster, containing either one or two THAs. Ranstam et al. (2011) suggested to apply the methodology of Hoffman et al. (2001), which is valid for data with clusters of nonignorable size, meaning that the risk for the outcome is related to the cluster size. For within-cluster resampling, a large number of data sets is created by randomly selecting one observation per individual. The estimator is computed on each data set, after which all estimates are averaged, resulting in the within-cluster resampling estimator.

Hoffman et al. (2001) prove asymptotic normality in the context of generalized linear models, and the main proof concept can be adapted to the competing risks setting, when combined with results in Cheng et al. (1998); Fine and Gray (1999) and Lin (1997). This extension would require assuming that both hips follow the same model, which seems unlikely to be true. The resulting estimator would represent the cumulative incidence of revision for a randomly sampled hip from a randomly sampled patient, and again, it is not clear how meaningful this would be in practice.

The within-cluster resampling procedure is reminiscent of the block or cluster bootstrap, but these methods differ in execution and aim. Suppose we have observed C clusters. The resampled datasets of the cluster bootstrap arise by sampling C clusters with replacement (Davison and Hinkley, 1997), while for within-cluster resampling, exactly one observation is sampled from each cluster. In the arthroplasty example, the cluster bootstrap would be performed by sampling the patients with replacement, while within-cluster sampling proceeds by sampling one hip per patient.

Regarding the difference in aim, the cluster bootstrap is intended to find the sampling distribution of the estimator, which would in our example be the variance of the estimated cumulative incidence of revision for a randomly sampled hip from the population. The two methods coincide only when there is no correlation between units in a cluster.

Shared gamma frailty model

A shared gamma frailty model was proposed to model the within-patient correlation (Ripatti and Palmgren, 2000), and has been applied since (Robertsson and Ranstam, 2003; Schwarzer et al., 2001). A disadvantage of these models is that the correlation is explicitly modeled, and the underlying assumptions do not necessarily hold for arthroplasty data. In particular, only positive correlation between the two THAs can be induced (Wienke, 2003). Not much is known about the correlation between two THAs in one patient. A positive one is possible, e.g. if the patient is very active, both prostheses are prone to earlier failure. However, two prostheses in one patient can be negatively correlated. If the patient favors one of the prostheses, then the prosthesis bearing the most stress is likely

to fail early while the other prosthesis is likely to survive longer. Thus, the shared gamma frailty model does not seem to be entirely adequate.

Cluster Fine-Gray

An extension of the Fine-Gray proportional subdistribution hazards model for clustered data is available (Zhou et al., 2012). More details on standard Fine-Gray regression can be found in Section 6.2. The cluster version has, to the best of our knowledge, not been applied to arthroplasty data yet. The cumulative incidence is estimated using standard Fine-Gray methodology under an independence working assumption, after which the variance is estimated using a sandwich variance estimator. The method was designed for settings where there are unobserved shared factors across individuals, such as multicenter trials or family studies. The correlation structure remains unspecified, making this method more attractive than a frailty model for arthroplasty data.

6.3.2 Methods for the time-dependent bilaterality status

In the terminology of Kalbfleisch and Prentice (2002), a patient's bilaterality status can be viewed as an *internal* time-dependent covariate, meaning that the possibility of its observation depends on the survival status of the patient. Internal time-dependent covariates pose a challenge in competing risks analysis, as their very observation at some time point t informs us that the probability of survival up until time t conditional on the time-dependent covariate is equal to one. It is possible to estimate cause-specific hazards, but prediction of cumulative incidences is not possible when an internal time-dependent covariate is included (Andersen et al., 1993; Cortese and Andersen, 2009). This makes the method of Lie et al. (2004), who propose to include a time-dependent covariate that contains information on a patient's bilaterality status and revision status of the opposite hip, unsuitable for our purposes.

If one's goal is to study the entire patient population, without any specific interest in the bilateral patients, the time-dependence problem can be avoided by only including each patient's first THA in the analysis, as discussed in Section 6.3.1. In this section, we discuss methods for the situation where the goal is to study bilateral patients specifically, or when the loss of data resulting from excluding the second limb is considered prohibitive.

Cortese and Andersen (2009) discuss three methods to incorporate a time-dependent covariate: a multistate model with additional transient states, the landmark analysis of Van Houwelingen (2007), or an extended competing risks model in which all possible combinations between the levels of the time-dependent covariate and cause-specific events are included as final states. These alternatives require a change of research question: the multistate model takes the per-patient point of view as opposed to the per-hip point of view, landmark analysis yields estimates conditional on event-free survival up until a landmark time, as does the extended competing risks model. Before reviewing these three options, we discuss the potential for immortal time bias.

Immortal time bias

A basic principle in survival analysis is that subgroups defined by patient characteristics that are not known at the start of follow-up (such as receiving a second THA), can only be compared with the greatest caution. The reason is the immortal time bias, a well-known phenomenon in observational studies, resulting from flawed statistical analysis (Lévesque et al., 2010; Suissa, 2007). Immortal time refers to a period of follow-up during which the study outcome, which may be death or another event (e.g. revision surgery), cannot occur. It was first described in the context of heart transplant data, when it was noted that the observed improved survival of heart transplant patients was due to selection bias: only patients who survived long enough to receive a heart transplant were included in the transplant group (Gail, 1972).

Analyses of arthroplasty data risk being affected by the immortal time bias as well. The immortal time bias arises when patients with staged bilateral THA are studied as a separate subgroup, because only those patients who survive long enough to be able to receive the second implant are observed. The bias occurs both when revision of one of the implants or death are taken as the endpoints. With arthroplasty data, when the outcome of interest is revision, the bias is subtle. Revision of the first hip does not prevent a patient from joining the staged bilateral group, and thus there is no obvious immortal time bias. However, there is the competing risk of death.

The underlying mechanism of the immortal time bias is illustrated through an artificial example, in which 50% of patients will become staged bilateral exactly 2 years after their index surgery. The first-placed implants of unilateral and bilateral patients are compared. The implants of all patients behave the same: they have a 30% probability of revision after exactly 3 years. In addition, each patient has a 20% probability of dying after 1 year. All percentages are chosen for illustrative purposes and are not meant to be realistic. We assume independence for all events. The process is visualized in Figure 6.1.

When the unilateral and staged bilateral subgroups are created at the end of follow-up, patients that would have become staged bilateral at the 2-year mark but died before realizing that potential, are observed to be unilateral. This leads to an estimate of a zero probability of death for staged bilateral patients, while the cumulative incidence of death is overestimated for unilateral patients. The reverse happens for revision: the cumulative incidence of revision is overestimated for staged bilateral patients, as the competing risk of death is not observed, while it is underestimated for unilateral patients, because the risk set is made artificially large by the inclusion of patients who would have become staged bilateral if they had not died before the second surgery could take place.

The severity of the effect of the immortal time bias depends on the revision, mortality and bilaterality rates, and also on the research question. With 5% revision, 5% mortality and 20% bilaterality, the bias in the artificial data example is inconsequential for the cumulative incidence of revision, but still relatively large for death. Moreover, statistically significant differences in implant survival between two groups can be very small, even less than 1%, when the follow up is short. In such a case, even a small bias may be large enough to give the false impression of a difference between subgroups where there is actually none. In addition, the Swedish Hip Arthroplasty Register reports 23-year revision rates of up to 38.5% for men who are 50-59 years at index surgery (SHAR, 2014). After such a long follow up, immortal time bias may significantly affect analyses, and thus clini-

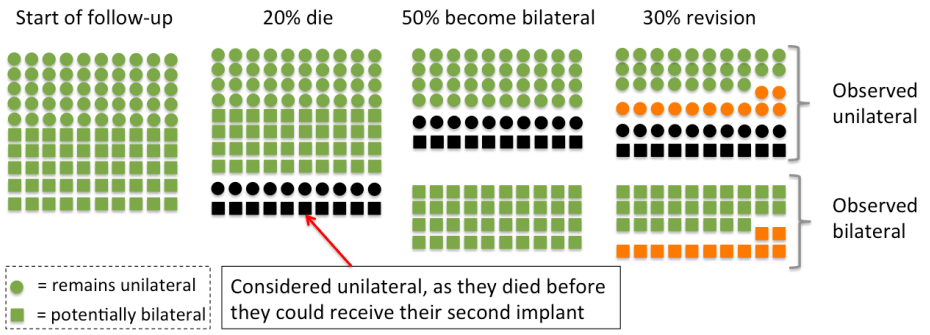

Figure 6.1: Visualization of the artificial data example. Circles denote patients who will only have one implant, while squares indicate patients who will become bilateral at the 2-year mark. Green indicates event-free patients, black patients who die before experiencing revision, and orange patients whose prosthesis has been revised. The subgroup analysis ignores the fact that some patients will have died before realizing their potential of becoming bilateral, and thus some potentially bilateral patients will be considered unilateral.

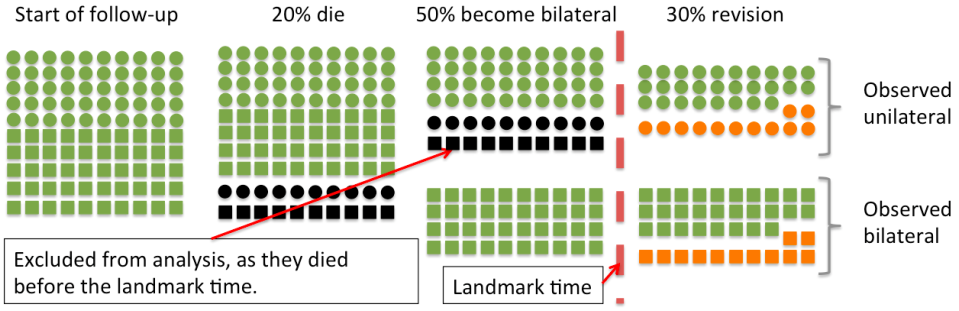

Figure 6.2: A landmark time is chosen, in this case after the patients become bilateral. All patients who died or were revised before the landmark time are excluded from the analysis.

cal results based on subgroup analysis with staged bilateral patients should be interpreted with caution.

Landmark analysis

Landmark analysis allows for comparison of unilateral and staged bilateral patients without the risk of immortal time bias (Cortese and Andersen, 2009; Van Houwelingen, 2007). The first step is to choose a landmark time of for example 2 years. The choice of landmark time should be guided by the research question. Only patients who are still alive and have not experienced revision yet at the landmark time are included in the analysis. This ensures a fair comparison between the two groups, as both need to survive for the same minimum amount of time in order to be included in the analysis.

The next step is to create the subgroups: patients who have become bilateral by the landmark time, and patients who were unilateral at the landmark time. As only each patient's status at the landmark time is considered, the latter group includes patients who may receive a second implant after the landmark time. The procedure is illustrated in Figure 6.2.

When the landmark subgroups have been made, the cumulative incidence can be estimated, for example by using one of the methods described in Section 6.2. The interpretation of the resulting models is conditional on the landmark time. Thus, conclusions can be drawn for comparison of unilateral and staged bilateral patients, conditional on the fact that these patients were still alive and did not undergo revision by the landmark time. This is a limitation to the method: the conclusions only hold for patients who are still alive and unrevised by the landmark time point. This is not a negative per se, as this question will be of interest to a patient who has survived some time unrevised since the primary or index THA. However, excluding the first few postoperative months or years from analysis may not be satisfactory in a situation where mortality or revision risk are especially high immediately following surgery.

Extended competing risks model and multistate models

The second approach discussed by Cortese and Andersen (2009) is an extended competing risks model, which has all possible combinations between internal covariate levels and cause-specific events as final states. In the case of arthroplasty data, such a model could be represented as in Figure 6.3.

The change in status from unilateral to bilateral comes bundled with the introduction of a second outcome: revision of the second hip. Thus, the outcome "revision" needs a more precise definition, such as "revision of the first THA", in which case the outcome of the second THA is disregarded.

The disease process of a patient can be more fully captured by a multistate model with transient states. Such a multistate model allows inclusion of a patient's second THA in a natural manner. Another advantage of these extended models is that they allow us to take a per-patient point of view, which is more useful to the orthopaedic surgeon than the classical per-hip point of view. See Figure 6.4 for an example of such a model.

There is a Markovian assumption behind this model, which can be relaxed. It may be the case that the probability of transitioning from for example "bilateral" to one of the

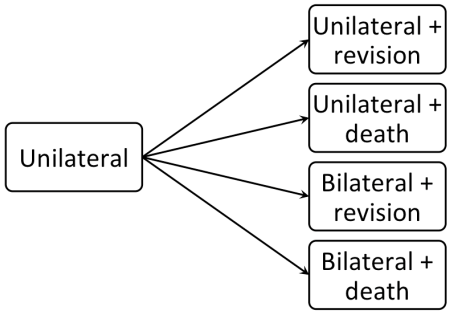

Figure 6.3: Extended competing risks model for Total Hip Arthroplasty.

revised states, depends on the amount of time spent as a unilateral patient. It is possible to model the intensity regulating the transitions as a function of the time spent as a unilateral patient, resulting in a semi-Markov model (Cortese and Andersen, 2009; Putter et al., 2007).

While multistate models have been rarely used in orthopaedic studies, there has been a successful application to the data of the Australian Orthopaedic Association National Joint Replacement Registry (Gillam et al., 2013, 2012).

6.3.3 Clinical relevance

Two reviews of arthroplasty studies found that is commonly believed that the bilateral patients do not affect the results of the analyses too much, and thus the dependence of their observations is often ignored (Bryant et al., 2006; Ranstam et al., 2011). Robertsson and Ranstam (2003) find that the effect of subject dependency in total knee arthroplasty is negligible, and explain this by saying that the source of the bias generated by ignoring dependency consists solely of bilateral patients with revisions on both sides, of which there are very few. Findings of Ripatti and Palmgren (2000), Schwarzer et al. (2001), Visuri et al. (2002) and Lie et al. (2004) for THA are similar. A contributing factor is that hip implant survival is very high.

The findings that ignoring the within-patient dependence does not significantly affect results are all within the context of questions about the entire patient population. Whether ignoring the presence of bilateral patients is problematic depends on the goal of the analysis, and on the similarity of the outcomes for the two prostheses. If one is interested in the time to revision for any hip, then ignoring the dependence may be a pragmatic solution if the first and second THAs have similar survival properties and similar associations with the covariates, and especially if implant survival is high in general. In that case, the ignored dependence will likely only affect the confidence intervals. However, if the implants of bilateral patients have different survival properties than unilateral prostheses, grouping everyone together without extra consideration does not make much sense. In that cases, studying unilateral and bilateral patients separately will provide more useful clinical insights.

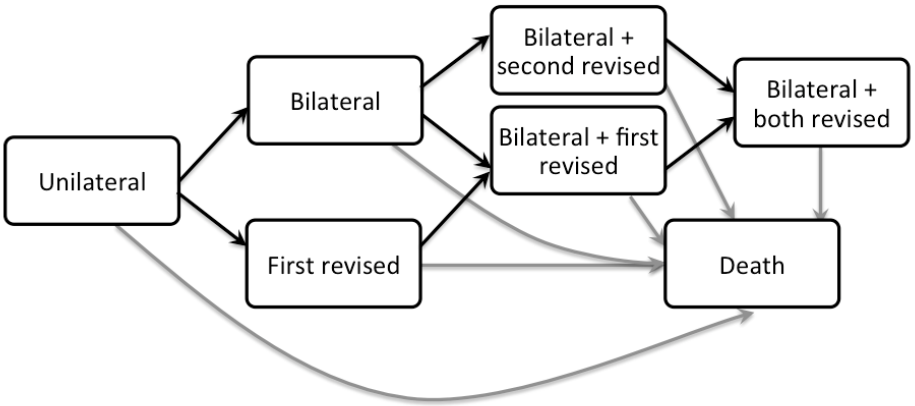

Figure 6.4: Multistate model for Total Hip Arthroplasty.

When the research question is concerned with bilateral patients, caution is warranted. The main potential pitfalls lie in the time-dependence of a patient's bilateral status. Extra care needs to be taken when the proportion of bilateral patients is high, and when the patients who become staged bilateral tend to do so relatively long after the first surgery. A naive subgroup analysis may be affected by immortal time bias. Landmark analysis or a multistate model seem appropriate solutions in this case.

In any analysis of arthroplasty registry data, researchers should carefully consider the impact the bilateral patients may have on their results, define their research population precisely, and select the statistical method accordingly.

6.4 Data structure

The data set contains data on 161,434 primary total hip arthoplasties, undertaken between 2007 and 2014. Arthroplasties after tumors or fractures, and hemiarthroplasties were not included. The survival information is captured in the following four variables:

1. **Status_revision**: indicates whether the hip was revised.
2. **Status_death**: indicates whether the patient has died.
3. **Surv_revision**: time at risk until revision.
4. **Surv_death**: time at risk until death.

In order to illustrate the time to event structure of the data, consider the following patients (only status indicators and time at risk shown for clarity):

Patient	Status_revision	Status_death	Surv_revision	Surv_death
103867	0	0	7.67	7.67
99702	1	0	3.97	6.56
88945	0	1	0.18	0.18
6645	1	1	0.13	0.52

Patient 103867 was under follow up for 7.67 years and was still alive at the end of follow up, without revision of his or her implant. Patient 99702 was under follow up for 6.56 years. After 3.97 years, his or her hip implant was revised. The patient was still alive at the end of follow up. Patient 88945 died after 0.18 years of follow up, without revision of his or her implant. The implant of patient 6645 was revised after 0.13 years, and the patient died 0.39 years later, at 0.52 years of follow up.

Each line in the data set corresponds to one hip. However, there are bilateral patients included in the data set, with a hip implant on both sides. Some examples in the data:

Patient	Status_revision	Status_death	Surv_revision	Surv_death
5	0	0	3.16	3.16
5	0	0	2.30	2.30
3044	0	1	1.63	1.63
3044	1	1	0.15	1.13
22112	0	0	4.47	4.47
22112	1	0	1.36	2.86

Patient 5 received a second hip implant 0.86 years after the first, and was then followed for another 2.30 years. During that time, none of the implants were revised, and the patient was still alive at the end of follow up. Patient 3044 received his or her second implant after 0.5 years. The second implant was revised 0.15 years after its placement. The first was never revised. The patient died 1.63 years after the first prosthesis was implanted. Patient 22112 received his or her second implant after 1.61 years, and it was revised 1.36 years later. The patient was still alive, without revision of the first implant, at the end of follow up at 4.47 years.

The statistical complications associated with the presence of bilateral patients in the data set are discussed in Section 6.3. Before proceeding to the data analysis in Section 6.5, we describe the remaining variables in the data set. The variables used in the model are listed below.

1. **Age**: age of patient at index surgery.
   Converted to the five age categories used by the LROI: younger than 50, 50-59, 60-69, 70-79, 80 and older.

2. **GENDER**: gender of patient.

3. **ASACLASH**: American Society of Anesthesiologists (ASA) classification.
   1: A normal healthy patient. 2: A patient with mild systemic disease. 3: A patient with severy systemic disease. 4: A patient with severe systemic disease that is a constant threat to life (ASA, 2014).

4. **DIAGH**: diagnosis.
   The nine diagnoses in the data set were combined into five diagnostic groups, following the recommendation of the clinician: osteoarthritis, post-Perthes and dysplasia, rheumatoid and inflammatory arthritis, osteonecrosis, and late posttraumatic combined with all other diagnoses.

5. **FIXH_incl_rev**: type of fixation of the hip implant.
   Cementless, hybrid, cemented or reversed hybrid.

6. **Hospitaltype**: type of hospital.
   General, academic, or private. Taken as a proxy for unmeasured confounders, outcome not reported.

The distribution of the patient characteristics is shown in Figure 6.5.

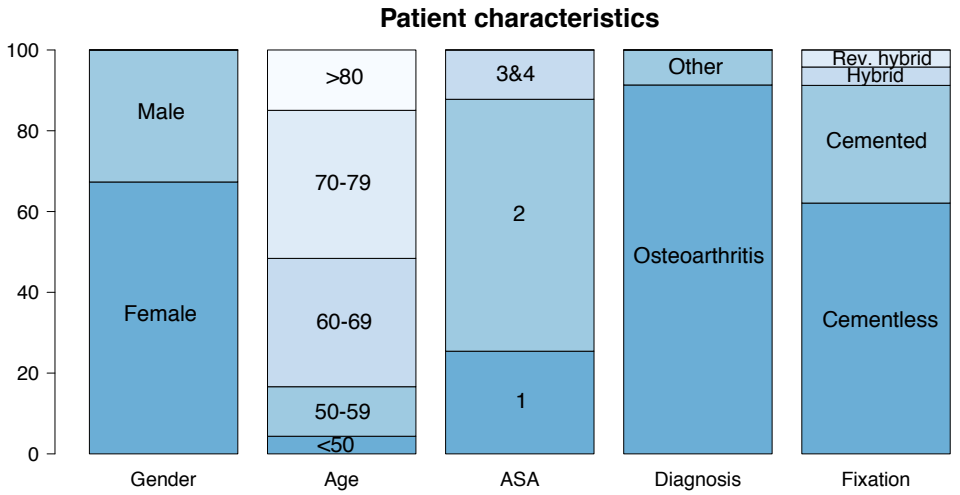

Figure 6.5: Barplot of patient characteristics in the LROI data set.

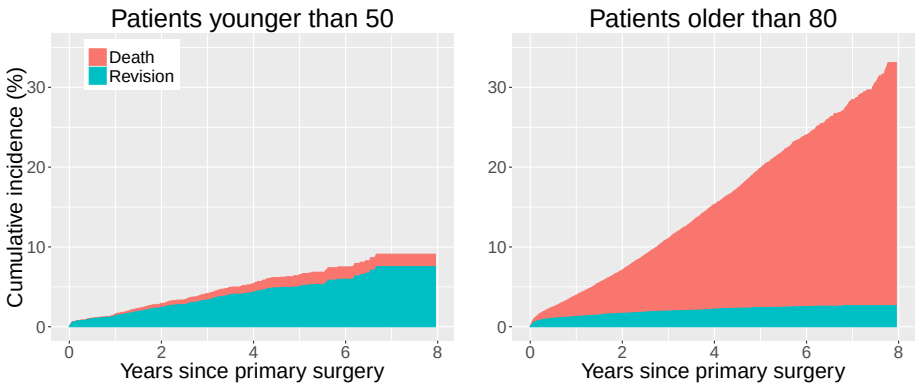

Figure 6.6: Aalen-Johansen estimator of the cumulative incidences of revision and death, for the youngest and oldest patients. Estimated using each patient's first THA.

Besides those variables, the data set contained information on the side of operation, year of operation and revision, and whether the hip had been operated on before. These were left out of the model because of redundancy or, in case of previous operations on the same hip, low completeness.

6.5 Results on the LROI data

This section contains results on data from the Dutch Arthroplasty Register (LROI). The analysis is based on total hip arthroplasty (THA) and subsequent revision surgeries performed on a total of 161,434 hips in 144,513 patients, between January 1st, 2007 and December 31st, 2014. We aim to identify variables associated with the cumulative incidence of revision, for all patients and in particular the bilateral patients.

A limitation of these analyses is that it requires identifying whether a THA was a patient's first or second. This is not problematic for patients whose both surgeries took place after 2007, but it is for bilateral patients whose first THA took place before the establishment of the LROI. If a patient's first THA happened before 2007, and his or her second THA happened after 2007, then only the second THA is recorded in the data set, and we would require an indicator to alert us to the fact that it is that patient's second, not first THA. Such an indicator exists in the form of the Charnley score, but this score has only been recorded sinds mid 2013, and we do not have access to it at time of writing. This is discussed in more detail in 6.5.3.

We first describe the analyses and present the results. We discuss the results, limitations of these analyses and plans for future analyses in Section 6.5.5.

6.5.1 Competing risks

The need for competing risks methods is illustrated in Figure 6.6. It shows the cumulative incidences of revision and death, estimated separately for the youngest and the oldest patients. The cumulative incidences were estimated using each patient's first THA, noting that for some patients, this will actually be their second, as explained above.

The Figure shows a very strong competing risk of death for patients older than 80. For patients under 50, the competing risk of death in this relatively short amount of follow up is so small as to be negligible. Given the average age of patients undergoing THA, which is 69, the competing risk of death cannot be ignored.

6.5.2 All patients

Before zooming in on the bilateral patients, we consider the entire patient population. Following the recommendations of Grambauer et al. (2010) and Latouche et al. (2013), both Fine-Gray regression and cause-specific Cox regression are performed.

We use Fine-Gray regression for clustered data (Zhou et al., 2012) to find variables associated with revision. As discussed in Section 6.3, this analysis does not account for the time between two THAs for bilateral patients. As a form of sensitivity analysis, standard Fine and Gray regression was also performed on the entire data set, and on all first THAs

Table 6.1: Fine-Gray regression for all patients. Reference category between parentheses.

Variable	Cluster Fine-Gray all data		Fine-Gray all data		Fine-Gray first THAs	
	coefficient	s.e.	coefficient	s.e.	coefficient	s.e.
Gender (female)						
Male	0.080	0.036	0.080	0.036	0.076	0.037
Age (< 50)						
50-59	-0.073	0.083	-0.073	0.082	-0.084	0.087
60-69	-0.285	0.080	-0.285	0.079	-0.277	0.083
70-79	-0.300	0.083	-0.300	0.082	-0.312	0.086
≥ 80	-0.426	0.094	-0.426	0.093	-0.407	0.098
ASA (ASA 1)						
ASA 2	0.090	0.040	0.090	0.040	0.086	0.042
ASA 3 & 4	0.234	0.060	0.234	0.060	0.249	0.062
Diagnosis (Osteoarthritis)						
Osteonecrosis	0.027	0.097	0.027	0.096	0.063	0.098
Post-Perthes/Dysplasia	-0.123	0.110	-0.123	0.110	-0.135	0.115
Late posttraumatic	0.434	0.091	0.434	0.091	0.421	0.093
Rheum./infl. arthritis	-0.085	0.162	-0.085	0.163	-0.140	0.177
Fixation (Cementless)						
Cemented	-0.559	0.047	-0.559	0.046	-0.533	0.049
Hybrid	-0.260	0.089	-0.260	0.088	-0.284	0.094
Reversed hybrid	0.046	0.078	0.046	0.078	0.081	0.081

(which will in some cases be the second THA, as explained above). The results are given in Table 6.1.

Cause-specific Cox regression was done for revision and death, both on the first THAs and on all THAs. The results are given in Table 6.2.

The subdistribution hazard ratios resulting from cluster Fine-Gray, as well as the hazard ratios resulting from cause-specific Cox regression on all patients are given in Table 6.3, together with the p-values and the numbers of revisions and deaths.

The cumulative incidence of revision is associated with gender, age, ASA score, diagnosis and type of fixation. Men are more likely to experience revision than women. The cumulative incidence of revision decreases with age, and increases with ASA score. It is less for hybrid fixation and even smaller for cemented fixation, compared to cementless fixation. This is set in context and discussed in Section 6.5.5.

As a visual check of the proportionality assumptions for Fine-Gray regression and cause-specific Cox regression, nonparametric estimates of the cumulative incidences of revision and the cause-specific cumulative hazards of revision and death are given in Figures 6.7, 6.8 and 6.9.

Table 6.2: Cause-specfic Cox regression for all patients. Reference category between parentheses.

Variable	First THAs				All data			
	Revision		Death		Revision		Death	
	coefficient	s.e.	coefficient	s.e.	coefficient	s.e.	coefficient	s.e.
Gender (female)								
Male	0.085	0.037	0.478	0.027	0.088	0.035	0.476	0.026
Age (< 50)								
50-59	-0.079	0.086	0.687	0.159	-0.068	0.082	0.767	0.155
60-69	-0.270	0.082	1.099	0.150	-0.278	0.078	1.162	0.147
70-79	-0.296	0.085	1.803	0.149	-0.284	0.081	1.866	0.146
≥ 80	-0.365	0.096	2.583	0.149	-0.384	0.092	2.636	0.147
ASA (ASA 1)								
ASA 2	0.091	0.042	0.313	0.039	0.095	0.040	0.315	0.038
ASA 3 & 4	0.281	0.062	1.100	0.044	0.265	0.059	1.112	0.042
Diagnosis (Osteoarthritis)								
Osteonecrosis	0.078	0.096	0.672	0.061	0.040	0.094	0.651	0.060
Post-Perthes/Dysplasia	-0.132	0.115	0.023	0.120	-0.120	0.110	0.049	0.115
Late posttraumatic	0.441	0.092	0.645	0.062	0.453	0.091	0.636	0.062
Rheum./infl. arthritis	-0.132	0.176	0.499	0.106	-0.078	0.163	0.484	0.103
Fixation (Cementless)								
Cemented	-0.530	0.048	0.138	0.029	-0.556	0.046	0.140	0.028
Hybrid	-0.282	0.093	0.195	0.055	-0.258	0.088	0.200	0.053
Reversed hybrid	0.082	0.081	0.048	0.078	0.048	0.078	0.060	0.075

Table 6.3: Numbers of events (all patients). Subdistribution hazard ratios (SHR) from cluster Fine-Gray, and hazard ratios (HR) from cause-specific Cox on all patients. Reference category between parentheses.

Variable				Cluster Fine-Gray		Cause-specific Cox			
				revision		revision		death	
	event-free	revisions	deaths	SHR	p-value.	HR	p-value	HR	p-value
Gender (female)	101,121	2,482	4,410	1		1		1	
Male	48,305	1,395	2,734	1.08	0.025	1.09	0.013	1.61	< 0.001
Age (< 50)	6,729	252	55	1		1		1	
50-59	18,816	641	297	0.93	0.38	0.93	0.41	2.15	< 0.001
60-69	48,796	1,256	1,144	0.75	< 0.001	0.76	< 0.001	3.20	< 0.001
70-79	54,819	1,293	2,949	0.74	< 0.001	0.75	< 0.001	6.46	< 0.001
≥ 80	20,894	448	2,724	0.65	< 0.001	0.68	< 0.001	13.96	< 0.001
ASA (ASA 1)	37,105	999	946	1		1		1	
ASA 2	90,014	2,151	3,635	1.09	0.025	1.10	0.017	1.37	< 0.001
ASA 3 & 4	16,371	469	1,910	1.26	< 0.001	1.30	< 0.001	3.04	< 0.001
Diagnosis (Osteoarthritis)	137,592	3,491	6,362	1		1		1	
Osteonecrosis	4,316	136	328	1.03	0.78	1.04	0.67	1.92	< 0.001
Post-Perthes/Dysplasia	3,628	94	84	0.88	0.26	0.89	0.27	1.05	0.67
Late posttraumatic	3,304	134	298	1.54	< 0.001	1.57	< 0.001	1.89	< 0.001
Rheum./infl. arthritis	1,522	41	104	0.92	0.60	0.92	0.63	1.62	< 0.001
Fixation (Cementless)	93,192	2,718	3,254	1		1		1	
Cemented	42,659	752	3,124	0.57	< 0.001	0.57	< 0.001	1.15	< 0.001
Hybrid	6,668	151	448	0.77	0.0033	0.77	0.003	1.22	< 0.001
Reversed hybrid	6,331	192	205	1.05	0.56	1.05	0.54	1.06	0.42

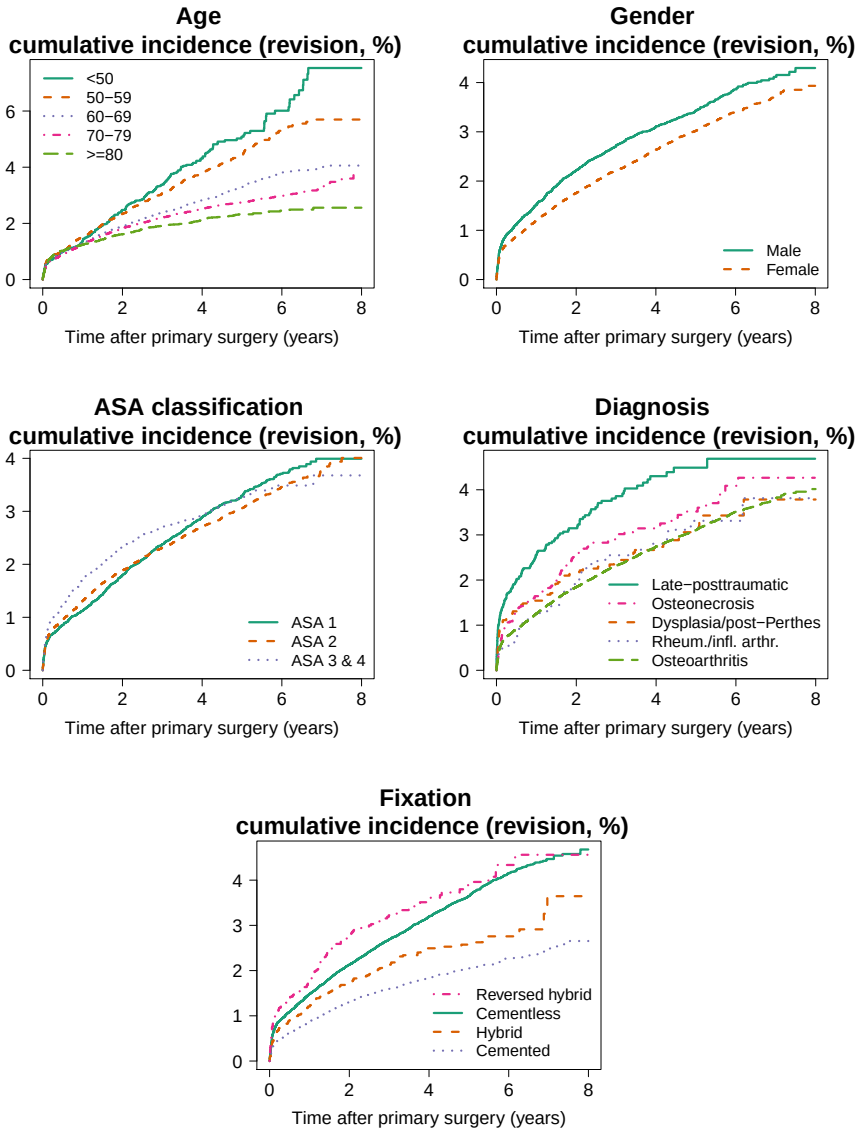

Figure 6.7: Aalen-Johansen estimates of the cumulative incidences of revision, using all hips in the data set.

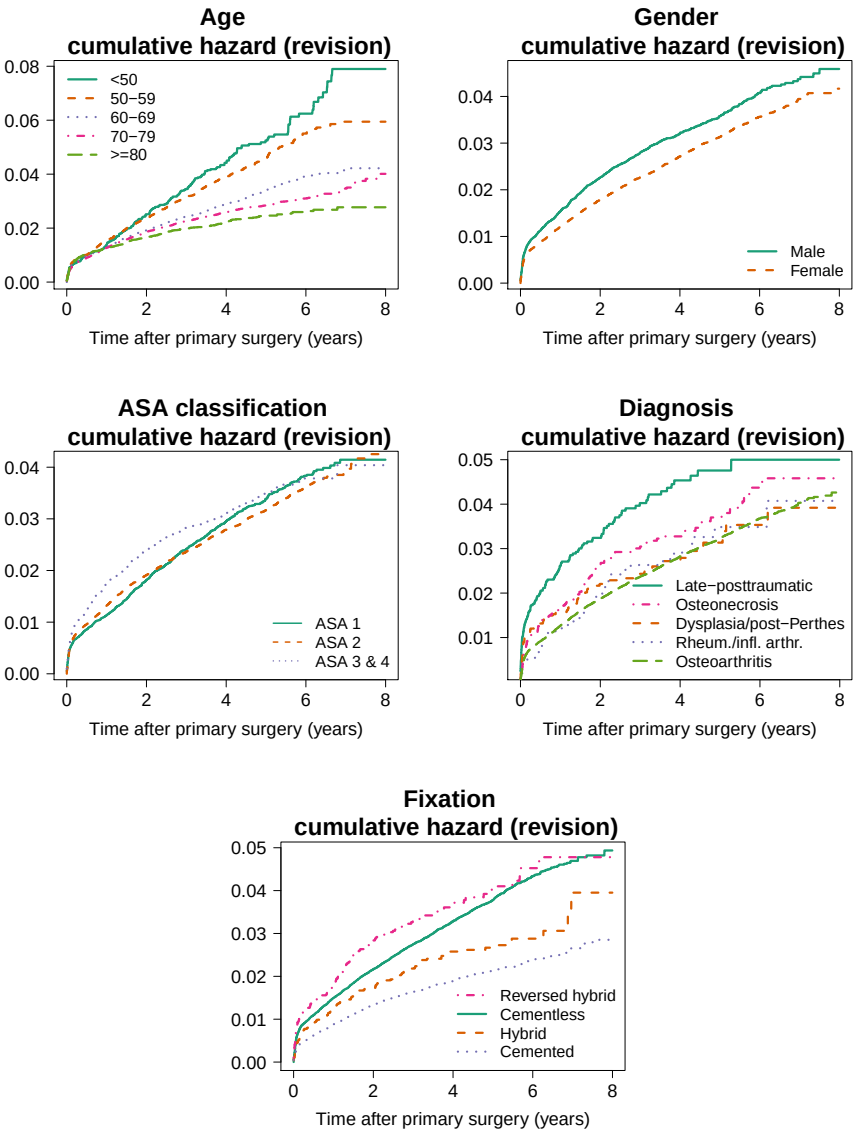

Figure 6.8: Nelson-Aalen estimates of the cumulative hazard of revision, using all hips in the data set.

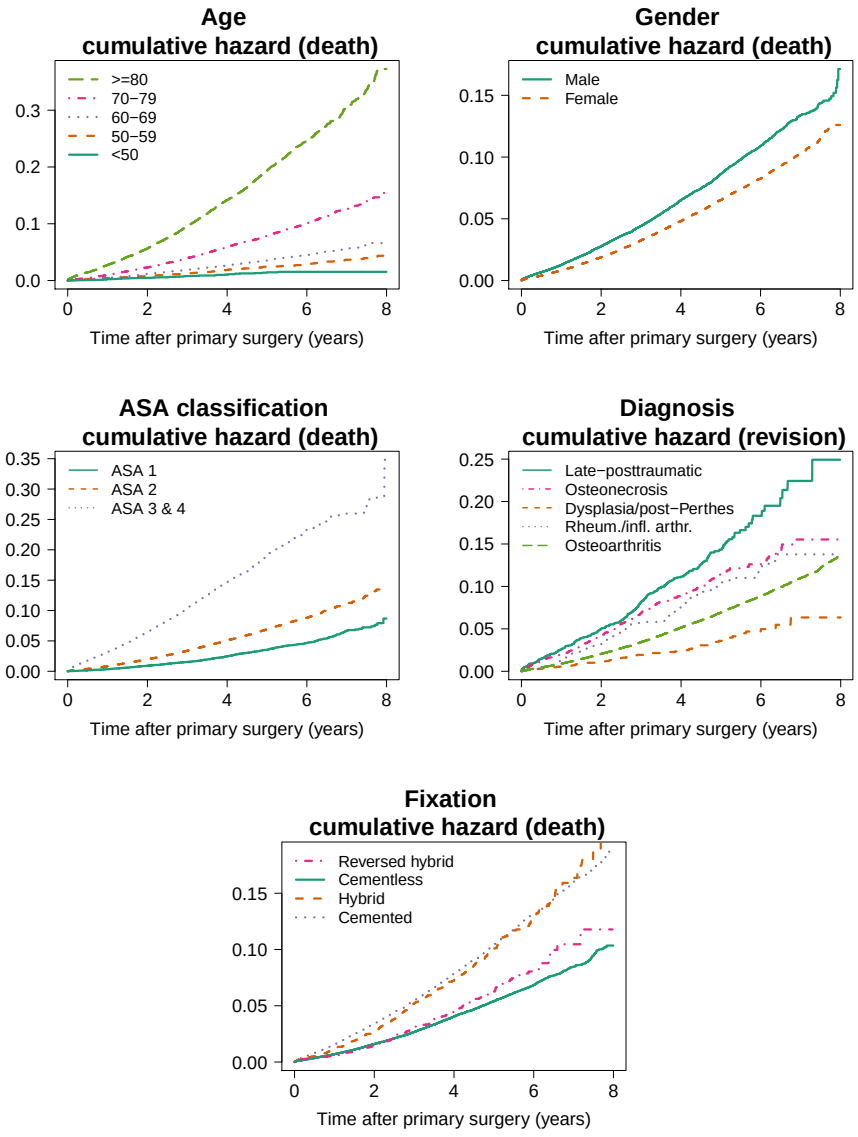

Figure 6.9: Nelson-Aalen estimates of the cumulative hazard of death, using all hips in the data set.

Table 6.4: Number of known second THAs in the LROI data.

	2007	2008	2009	2010	2011	2012	2013	2014
Number of THAs (total)	7,938	13,781	19,976	21,992	22,713	23,971	24,631	26,438
Number of second THAs	197	632	1,304	1,975	2,539	2,964	3,436	3,874
Percentage second THAs	2.5%	4.6%	6.5%	9.0%	11.2%	12.4%	13.9%	14.7%

6.5.3 Comparison of unilateral and bilateral patients

We compare rates of revision for the first implanted hip for unilateral and staged bilateral patients at the landmark time of 1 year, meaning that we include those patients who had not yet experienced revision after 1 year and divide them into groups who have received either one or two prostheses by 1 year.

As explained above, this analysis is problematic, because the "unilateral" group will contain some second THAs from bilateral patients. To get a sense of how many of these second THAs we may miss, we compute for each year the number of THAs that are known to be the second of a bilateral patient, because the corresponding first THA took place in or after 2007. These numbers are given in Table 6.4. For reference, the Charnley score was recorded in 2014, and in that year, 20% of THAs concerned the placement of a second hip (LROI, 2014).

To mitigate the problem of the unidentified second THAs, we only study patients whose first (known) procedure took place in 2010 or later. Based on clinical experience, we perform landmark analysis at the 1 year landmark, for 4.5 years of follow-up. In total, 75,397 patients were included in the unilateral group, and 5,031 in the bilateral group. Gray's test detects a difference in cumulative incidence of revision between patients who are unilateral or bilateral 1 year after the first THA ($p = 0.003$). The estimated cumulative incidences are given in Figure 6.10. As shown in Figure 6.10, the first implanted prosthesis of a patient who has become bilateral at the one year mark is less likely to be revised compared to unilateral prostheses, if we compare patients who have not undergone revision and are still alive one year after the first THR.

6.5.4 Second-implanted hips

For a comparison of the second-implanted hips of staged bilateral patients, no time-dependent covariates are required, as their time point of origin is the time of the second primary THA. We thus compute the unadjusted and adjusted cumulative incidences without any further considerations. Characteristics of the bilateral patients are given in Table 6.5.

The results from the Fine-Gray regression are given in Table 6.6. When we consider the second-placed hips of staged bilateral patients, the amount of time between the two surgeries is a significant variable. The unadjusted cumulative incidence of revision of the second hip is significantly different for patients whose second hip was placed more than one year after the first one, compared to patients whose hips were both placed within one year ($p = 0.009$). This is illustrated in Figure 6.11, which shows that the unadjusted cumulative incidence of revision for patients whose surgeries take place more than one year apart is higher than for patients whose surgeries take place within one year.

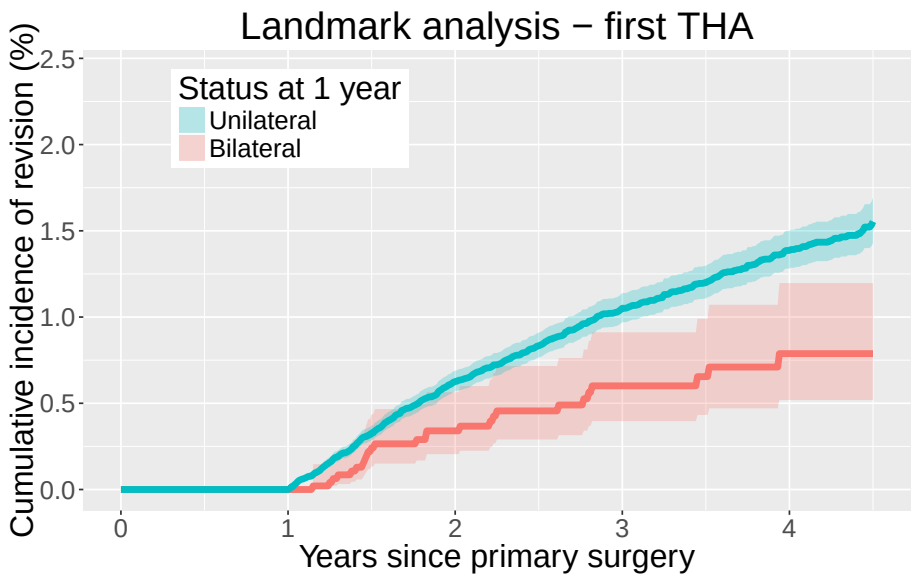

Figure 6.10: Aalen-Johansen estimator of the cumulative incidence of revision of the first hip implant for patients who are unilateral or bilateral and event-free at the 1 year landmark. The cumulative incidence of revision is higher for patients who are (still) unilateral 1 year after their first THA.

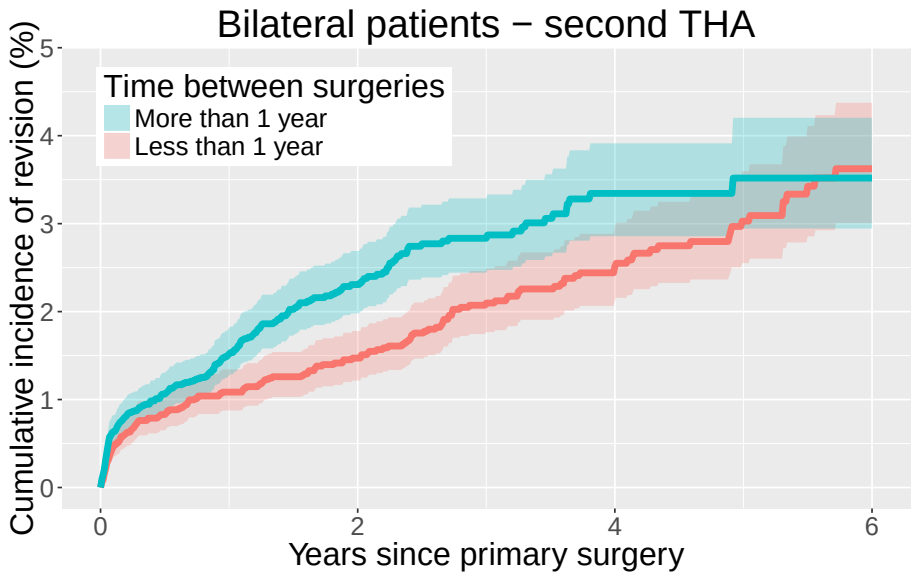

Figure 6.11: Aalen-Johansen estimator of the cumulative incidence of revision for a bilateral patient's second hip implant. The cumulative incidence of revision is higher for patients who receive their second THA after more than 1 year.

Table 6.5: Characteristics of bilateral patients as recorded at the time of the second THA, compared to all patients.

Characteristics	Second hip within 1 year	Second hip after more than 1 year	All patients
Total	8,027	8,894	161,434
Mean age	67.2	70.1	69.0
Female	68.7%	72.8%	67.3%
ASA 1	25.4%	20.9%	25.4%
ASA 2	64.2%	66.6%	62.4%
ASA 3&4	10.4%	12.5%	12.2%
Osteoarthritis	92.3%	95.6%	91.3%
Cementless fixation	64.1%	64.2%	62.1%

When the interoperative time is adjusted for, we find few remaining significant predictors for revision of the second prosthesis. Patients with cemented hips are less likely to experience revision than patients with cementless fixation, and the oldest patients are less likely to experience revision than the youngest. Gender, ASA and diagnosis do not appear to play a significant role when we consider only the second prostheses of bilateral patients.

6.5.5 Discussion and outlook

All patients

Outcome of Fine-Gray regression

The finding that young age, male sex, high ASA score, uncemented prostheses and an earlier trauma are risk factors for revision (Table 6.3) is consistent with previous studies (Prokopetz et al., 2012). Many of those previous analyses were done without accounting for the competing risk of death, but the conclusions still stand when it is corrected for. Explanations are available in the clinical literature. Younger patients are typically more active and heavier, leading to increased stress on the implant components compared to older patients (Johnsen et al., 2006). Higher mechanical stress may also explain the increased risk of revision for men compared to women, together with hip kinematics (Gallo et al., 2010). The ASA score is an indicator of a patient's preoperative health status, and can be predictive of the early functional status (Hooper et al., 2012). Regarding fixation, the lower cumulative incidence of revision for cemented implants compared to uncemented implants is well-documented (Makela et al., 2014). In addition, we find that hybrid prostheses have a lower risk of revision compared to uncemented prostheses, a finding for which previous studies found evidence in either direction (Prokopetz et al., 2012). Of the diagnoses included in this study, only those patients who receive a hip prosthesis long after a trauma have a significantly different risk of revision than patients who have a diagnosis of osteoarthritis. After a trauma, risk of dislocation is increased, as anatomic structures may be compromised (Mallory et al., 1999). Thus, the results are consistent with the clinical literature.

Table 6.6: Fine-Gray regression based on bilateral patients' second THA.

Variable	event-free	revisions	deaths	Subdistribution hazard ratio	95% CI	p-value
Time between surgeries (<1 year)	7,588	168	271	1		
≥ 1 year	8,463	198	233	1.37	1.10-1.70	0.004
Gender (female)	11,374	243	323	1		
Male	4,612	123	180	1.15	0.92-1.44	0.23
Age (< 50)	650	23	2	1		
50-59	1,973	69	29	0.96	0.57-1.62	0.87
60-69	5,444	118	96	0.64	0.38-1.09	0.096
70-79	5,962	132	224	0.74	0.44-1.26	0.27
≥ 80	1,999	24	152	0.44	0.23-0.84	0.013
ASA (ASA 1)	3,652	91	64	1		
ASA 2	10,316	229	277	1.12	0.87-1.45	0.37
ASA 3& 4	1,729	34	137	1.05	0.70-1.59	0.80
Diagnosis (Osteoarthritis)	15,100	341	470	1		
Osteonecrosis	392	7	16	0.66	0.30-1.47	0.31
Post-Perthes/Dysplasia	333	9	7	0.99	0.48-2.03	0.98
Late posttraumatic	58	3	4	2.91	0.92-9.18	0.069
Rheumatoid/inflammatory arthritis	168	6	7	1.35	0.60-3.07	0.47
Fixation (Cementless)	10,277	281	241	1		
Cemented	4,374	53	212	0.45	0.33-0.61	< 0.001
Hybrid	650	17	29	0.97	0.58-1.62	0.89
Reversed hybrid	671	12	16	0.73	0.41-1.30	0.28

Outcome of cause-specific Cox regression
The conclusions from the cause-specific Cox model are in line with the results from Fine-Gray regression. The subdistribution hazard ratios for revision and the cause-specific hazard ratios for revision are numerically very close (Table 6.3). This may be due to the heavy censoring (Grambauer et al., 2010): there were 161,434 hips in the data set and only 11,076 events (3,897 revisions and 7,179 deaths).

The added value of two separate analyses for death and revision is visible for those variables where the coefficients for the two hazards have opposite signs: old age, diagnoses of post-Perthes / dysplasia or rheumatoid/inflammatory arthritis, and cemented or hybrid fixation. In all these cases, the cause-specific hazard of revision is decreased; the cause-specific hazard of death increased, and the cumulative incidence of revision decreased. Besides the explanations already provided above, this analysis makes clear that another effect may be that patients with these characteristics are revised less frequently because the rate of occurence of death is increased.

The proportionality assumptions
We highlight two aspects of Figures 6.7, 6.8 and 6.9. First, all plots in Figure 6.7 are remarkably similar to the corresponding plots in Figure 6.8. We already observed that the subdistribution and cause-specific hazards for revision are numerically very close. This is most likely due to the heavy censoring.

The second aspect is that there is some evidence for violation of the proportional subdistribution/cause-specific hazards assumption. The assumption seems to hold for neither hazard for revision for the ASA score, age (first 2 years) and diagnosis. For age, the violation could be due to the categorization. The proportionality assumption does appear to be reasonable for gender and fixation, and for the cause-specific hazards of death. This can be investigated further using, for example, the methods listed in Section 6.2.

Sensitivity to the presence of bilateral patients
The differences between standard Fine-Gray and cluster Fine-Gray regression are negligible (Table 6.1). The estimated coefficients are the same, and the standard errors only differ on the third decimal place. The coefficients and standard errors estimated using only the first THAs are different compared to cluster Fine and Gray, but the signs of all coefficients are the same, and the same coefficients would be significant at the 5% level. The cause-specific Cox regression is not substantively impacted by the within-subject dependence of the bilateral patients either (Table 6.2); the differences between the estimated coefficients based on all THAs or only the first THAs are minimal, and conclusions based on a 0.05-cutoff for the p-values would be the same.

Bilateral patients

Results
Our results indicate that the cumulative incidence of revision is different for staged bilateral patients than for unilateral patients, and that staged bilateral patients are not a homogeneous subgroup. Interoperative time is an important factor to take into account. If a patient's second THR takes place within 1 year, not only does his or her first prosthesis

survive longer compared to unilaterally implanted prostheses, but his or her second prosthesis is less likely to be revised than the second prosthesis of a bilateral patient whose second THR took place more than 1 year after the first.

The results for the first bilateral implant compared to a unilateral implant correspond with the findings of the Swedish Hip Arthroplasty Register; they report better survival for the first bilateral THA compared to a unilateral implant (SHAR, 2014). However, in homogeneous subgroups consisting of patients with a diagnosis of osteoarthritis, no difference in survivorship of the first bilateral prosthesis compared to the unilaterals was found (Havelin et al., 1995; Lie et al., 2004; Visuri et al., 2002).

Most variables that were significant for all patients, are not significant at the 5% level for bilateral patients, when the time between surgeries is included as a categorical variable (Table 6.6). Only very old age and a cemented fixation remain significant. This may be in part because the time between surgery serves as a proxy for a patient's general health status and activity level, as will be discussed below.

Limitations

We must be careful not to draw causal conclusions, as the data are observational. Furthermore, there are limitations to the comparison of unilateral and bilateral patients. First of all, even after removing the data from 2007-2009, some second THAs will have been included in the "unilateral" group. Two studies indicate that the second THA has better survival than a unilateral implant, but the evidence is limited (Lie et al., 2004; Visuri et al., 2002). It is thus not clear how the presence of unidentified second THAs may have affected the estimates presented in Figure 6.10. A second limiting factor is that the landmark analysis precludes us from drawing conclusions about the risk of revision within the first year.

The analysis of the second THAs does not suffer from these limitations, and suggests that implant survival is better for patients who receive their second THA within 1 year after the first. When interpreting the results in Figure 6.11, the competing risk of death needs to be considered. A bilateral patient who receives his or her second implant after more than 1 year is on average older than a bilateral patient whose second surgery takes place within 1 year, as supported by Table 6.6. Being older, the patient may be at lower risk of revision. Yet Figure 6.11 and Table 6.5 indicate that patients who receive their second implant after more than 1 year have higher risk of revision, lending credence to the hypothesis that the two groups of bilateral patients differ from each other in some other respect.

Timing of the second THA

The protective effect of a shorter time between the two surgeries has been observed before (Havelin et al., 1995; Lie et al., 2004; Möllenhoff et al., 1994; Visuri et al., 2002). The cutoff for significant differences found in each of these studies has been different, and none of the studies accounted for the competing risk of death. The optimal lengths of interoperative time as reported by these studies are within 1 year (Visuri et al., 2002), within 2 years (Lie et al., 2004), or within 1-3 years (Möllenhoff et al., 1994).

Our results suggest that the relevant period may be as short as 1 year. However, again it must be stressed that these data are observational, and the conclusion that bilateral

THAs should be placed as soon as possible cannot be drawn.

We offer some clinical considerations on the observed protective effect of a shorter interoperative time period. One factor may be the relationship between activity levels and revision risk. A patient who receives two implants within 1 year may have other health issues associated with impeded mobility, thus putting less strain on the first replaced hip, leading to longer survival of the implant. Bilateral patients whose two surgeries are more than 1 year apart may have suffered from impaired mobility to a lesser extent, explaining why their implants are more prone to early failure than those of patients who received their second implant soon after the first.

On the other hand, with some diagnoses, patients may elect to have the second THA sooner rather than later. The patients who do so are likely to be in good health, and more satisfied with the outcome of the first THA. This may actually lead to worse survival of the implants, as these are generally more active patients.

A third factor may be that the group of osteoarthritis patients is not homogeneous, and that those who receive a second implant soon after the first represent a subgroup within the group of osteoarthritis patients for whom osteoarthritis should be considered a systemic disease.

Outlook

Only a randomized clinical trial can confirm hypotheses about interoperative time and improved outcomes for staged bilateral patients. The LROI is still relatively young. With the passing of time, more data will become available, allowing more detailed study of bilateral patients. A multistate model has been applied to the data from the Australian National Joint Replacement Registry, with promising results (Gillam et al., 2013, 2012). One insight from the Australian multistate model is that women are more likely than men to experience a second joint replacement surgery, which may be due to the lower mortality risk for women, or because women may have more extensive osteoarthritis. We expect that such a model, applied to the LROI data, would provide more insight into the path a patient may take from unilateral to possible bilateral, revision and/or death. The Dutch hip replacement data can be linked to knee replacement data, allowing for further study of patients with multiple implants.

Bibliography

Aalen, O. O. and Johansen, S. (1978), An empirical transition matrix for non-homogeneous Markov chains based on censored observations, *Scandinavian Journal of Statistics* **5**(3), 141–150.

Abbe, E., Bandeira, A. S. and Hall, G. (2014), Exact recovery in the stochastic block model. arXiv:1405.3267v4.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008), Mixed membership stochastic blockmodels, *Journal of Machine Learning Research* **9**, 1981–2014.

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov and F. Csaki, eds, 'Second International Symposium on Information Theory', Akademiai Kiado, Budapest, pp. 267–281.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993), *Statistical models based on counting processes*, Springer.

Andersen, P. K. and Pohar Perme, M. (2010), Pseudo-observations in survival analysis, *Statistical Methods in Medical Research* **19**(1), 71–99.

Andrews, D. F. and Mallows, C. L. (1974), Scale mixtures of normal distributions, *ournal of the Royal Statistical Society. Series B (Statistical Methodology)* pp. 99–102.

Andrews, M. and Baguley, T. (2012), Prior approval: the growth of Bayesian methods in psychology, *British Journal of Mathematical and Statistical Psychology* **66**(1), 1–7.

Armagan, A., Dunson, D. B. and Lee, J. (2013), Generalized double Pareto shrinkage, *Statistica Sinica* **23**, 119–143.

Armitage, P., McPherson, C. and Rowe, B. (1969), Repeated significance tests on accumulating data, *Journal of the Royal Statistical Society. Series A (General)* **132**(2), 235–244.

ASA (2014), ASA physical status classification system, https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system. [Online; version October 15, 2014].

Barndorff-Nielsen, O. (1978), *Information and exponential families in statistical theory*, Wiley.

Barron, A., Birgé, L. and Massart, P. (1999), Risk bounds for model selection via penalization, *Probability Theorie and Related Fields* **113**(3), 301–413.

Barron, A., Rissanen, J. and Yu, B. (1998), The minimum description length principle in coding and modeling, *IEEE Transactions on Information Theory* **44**(6), 2743–2760.

Belitser, E. (2014), On coverage and local radial rates of DDM-credible sets. arXiv:1407.5232.

Belitser, E. and Nurushev, N. (2015), Needles and straw in a haystack: empirical Bayes confidence for possibly sparse sequences. arXiv:511.01803.

Bem, D. J. (2011), Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect, *Journal of Personality and Social Psychology* **100**(3), 407–425.

Berger, J. and Wolpert, R. (1988), *The likelihood principle. Second edition*, Institute of Mathematical Statistics, Hayward, CA.

Beyersmann, J., Allignol, A. and Schumacher, M. (2012), *Competing risks and multistate models with R*, Springer.

Beyersmann, J. and Scheike, T. H. (2013), Classical regression models for competing risks, *in* J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim and T. H. Scheike, eds, 'Handbook of survival analysis', Chapman and Hall/CRC, pp. 157–177.

Beyersmann, J. and Schumacher, M. (2007), Letter to the Editor: Misspecified regression model for the subdistribution hazard of a competing risk, *Statistics in Medicine* **26**(7), 1649–1651.

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2015), The horseshoe+ estimator of ultra-sparse signals. arXiv:1502.00560v2.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2012), Bayesian shrinkage. arXiv:1212.6088.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2014), Dirichlet-Laplace priors for optimal shrinkage. arXiv:1401.5398.

Bickel, P. J. and Chen, A. (2009), A nonparametric view of network models and Newman-Girvan and other modularities, *Proceedings of the National Academy of Sciences of the United States of America* **106**(50), 21068–21073.

Bickel, P. J., Chen, A., Zhao, Y., Levina, E. and Zhu, J. (2015), Correction to the proof of consistency of community detection, *The Annals of Statistics* **43**(1), 462–466.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009), Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics* **37**(4), 1705–1732.

Bogdan, M., Ghosh, J. K. and Tokdar, S. T. (2008), A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing, *in* 'Beyond parametrics in interdisciplinary research: Festschrift in honor of professor Pranab K. Sen', The Institute of Mathematical Statistics.

Box, G. E. and Draper, N. R. (1987), *Empirical model-building and response surfaces*, Wiley.

Bryant, D., Havey, T. C., Roberts, R. and Guyatt, G. (2006), How many patients? How many limbs? Analysis of patients or limbs in the orthopaedic literature: a systematic review, *The Journal of Bone & Joint Surgery* **88-A**(1), 41–45.

Buchholz, H., Heinert, K. and Wargenau, M. (1985), Verlaufsbeobachtung von Hüftendoprothesen nach Abschluß realer Belastungsbedingungen von 10 Jahren, *Zeitschrift für Orthopädie* **123**, 815–820.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for high-dimensional data*, Springer-Verlag Berlin Heidelberg.

Bull, A. (2012), Honest adaptive confidence bands and self-similar functions, *Electronic Journal of Statistics* **6**, 1490–1516.

Burnham, K. and Anderson, D. (2004), Multimodel inference: Understanding AIC and BIC in model selection, *Sociological Methods & Research* **33**, 261–304.

Caron, F. and Doucet, A. (2008), Sparse Bayesian nonparametric regression, *in* 'Proceedings of the 25th International Conference on Machine Learning', ICML '08, ACM, New York, NY, USA, pp. 88–95.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2009), Handling sparsity via the horseshoe, *Journal of Machine Learning Research, W&CP* **5**, 73–80.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010), The horseshoe estimator for sparse signals, *Biometrika* **97**(2), 465–480.

Castillo, I. and Nickl, R. (2014), On the Bernstein von Mises phenomenon for nonparametric Bayes procedures, *Annals of Statistics* **42**(5), 1941–1969.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015), Bayesian linear regression with sparse priors, *The Annals of Statistics* **43**(5), 1986–2018.

Castillo, I. and van der Vaart, A. W. (2012), Needles and straw in a haystack: Posterior concentration for possibly sparse sequences, *The Annals of Statistics* **40**(4), 2069–2101.

Cavanaugh, J. E. (2012), [Catching up faster by switching sooner] : Discussion, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74**(3), 402–403.

Channarond, A., Daudin, J.-J. and Robin, S. (2012), Classification and estimation in the stochastic blockmodel based on the empirical degrees, *Electronic Journal of Statistics* **6**, 2574–2601.

Chen, K. and Lei, J. (2014), Network cross-validation for determining the number of communities in network data. arXiv:1411.1715v1.

Chen, Y. and Xu, J. (2014), Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. arXiv:1402.1267v2.

Cheng, S., Fine, J. P. and Wei, L. (1998), Prediction of cumulative incidence function under the proportional hazards model, *Biometrics* **54**(1), 219–228.

Chernoff, H. (1952), A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *The Annals of Mathematical Statistics* **23**(4), 493–507.

Clements, N., Sarkar, S. K. and Guo, W. (2012), Astronomical transient detection controlling the false discovery rate, *in* E. D. Feigelson and J. Babu, eds, 'Statistical challenges in modern astronomy V', Springer, pp. 383–396.

Côme, E. and Latouche, P. (2014), Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. arXiv:1303.2962.

Cortese, G. and Andersen, P. K. (2009), Competing risks and time-dependent covariates, *Biometrical Journal* **51**, 138–158.

Csardi, G. and Nepusz, T. (2006), The igraph software package for complex network research, *InterJournal Complex Systems* **1695**.

Csiszár, I. (1984), Sanov property, generalized I-projection and a conditional limit theorem, *The Annals of Probability* **12**(3), 768–793.

Damien, P., Wakefield, J. and Walker, S. (1999), Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**(2), 331–344.

Datta, J. and Ghosh, J. K. (2013), Asymptotic properties of Bayes risk for the horseshoe prior, *Bayesian Analysis* **8**(1), 111–132.

Davison, A. and Hinkley, D. (1997), *Bootstrap methods and their application*, Cambridge University Press.

Dawid, A. (1984), Present position and potential developments: Some personal views, statistical theory, the prequential approach, *Journal of the Royal Statistical Society. Series A (General)* **147**(2), 278–292.

Dienes, Z. (2011), Bayesian versus orthodox statistics: Which side are you on?, *Perspectives on Psychological Science* **6**(3), 274–290.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992), Maximum entropy and the nearly black object (with discussion), *Journal of the Royal Statistical Society. Series B (Methodological)* **54**(1), 41–81.

Edwards, W., Lindman, H. and Savage, L. J. (1963), Bayesian statistical inference for psychological research, *Psychological Review* **70**(3), 193–242.

Efron, B. (1986), How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association* **88**(394), 461–470.

Efron, B. (2008), Microarrays, empirical Bayes and the two-groups model, *Statistical Science* **23**(1), 1–22.

van Erven, T., Grünwald, P. D. and De Rooij, S. (2012), Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma (with discussion), *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74**(3), 361–417.

van Erven, T., Grünwald, P. and de Rooij, S. (2007), Catching up faster in Bayesian model selection and model averaging, *in* 'Advances in Neural Information Processing Systems', Vol. 20.

van Erven, T. and Harremoës, P. (2014), Rényi divergence and Kullback-Leibler divergence, *IEEE Transactions on Information Theory* **60**(7), 3797–3820.

Fine, J. P. and Gray, R. J. (1999), A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association* **94**(446), 496–509.

Forster, M. (2000), Key concepts in model selection: Performance and generalizability, *Journal of Mathematical Psychology* **44**, 205–231.

Gail, M. H. (1972), Does cardiac transplantation prolong life? A reassessment, *Annals of Internal Medicine* **76**(5), 815–817.

Gallo, J., Havranek, V., Zapletalova, J. and Lostak, J. (2010), Male gender, Charnley class C, and severity of bone defects predict the risk for aseptic loosening in the cup of ABG I hip arthroplasty, *BMC Musculoskeletal Disorders* **11**(1), 1–7.

Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2015), Achieving optimal misclassification proportion in stochastic block model. arXiv:1505.03772v5.

Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2016), Community detection in degree-corrected block models. arXiv:1607.06993.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), On asymptotically optimal confidence regions and tests for high-dimensional models, *The Annals of Statistics* **42**(3), 1166–1202.

van de Geer, S., Bühlmann, P. and Zhou, S. (2011), The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso), *Electronic Journal of Statistics* **5**, 688–749.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000), Convergence rates of posterior distributions, *The Annals of Statistics* **28**(2), 500–531.

Ghosal, S., Lember, J. and van der Vaart, A. (2008), Nonparametric Bayesian model selection and averaging, *Electronic Journal of Statistics* **2**, 63–89.

Ghosh, P. and Chakrabarti, A. (2015), Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors. arXiv:1412.8161v2.

Gillam, M. H., Lie, S. A., Salter, A., Furnes, O., Graves, S. E., Havelin, L. I. and Ryan, P. (2013), The progression of end-stage osteoarthritis: analysis of data from the Australian and Norwegian joint replacement registries using a multi-state model, *Osteoarthritis and Cartilage* **21**(3), 405–412.

Gillam, M. H., Ryan, P., Graves, S. E., Miller, L. N., de Steiger, R. N. and Salter, A. (2010), Competing risks survival analysis applied to data from the Australian Orthopaedic Association National Joint Replacement Registry, *Acta Orthopaedica* **81**(5), 548–555.

Gillam, M. H., Ryan, P., Salter, A. and Graves, S. E. (2012), Multi-state models and arthroplasty histories after unilateral total hip arthroplasties: Introducing the summary notation for arthroplasty histories, *Acta Orthopaedica* **83**(3), 220–226.

Giné, E. and Nickl, R. (2010), Confidence bands in density estimation, *The Annals of Statistics* **38**(2), 1122–1170.

Glover, F. (1989), Tabu search - part I, *ORSA Journal on Computing* **1**(3), 190–206.

Gradshteyn, I. S. and Ryzhik, I. M. (1965), *Table of integrals, series and products*, Academic Press.

Gramacy, R. B. (2014), *monomvn: Estimation for multivariate normal and Student-t data with monotone missingness*. R package version 1.9-5.

Grambauer, N., Schumacher, M. and Beyersmann, J. (2010), Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified, *Statistics in Medicine* **29**, 875–884.

Graves, S. (2010), The value of arthroplasty registry data, *Acta Orthopaedica* **81**, 8–9.

Gray, R. J. (1988), A class of k-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics* **16**(3), 1141–1154.

Griffin, J. E. and Brown, P. J. (2005), Alternative prior distributions for variable selection with very many more variables than observations, *Technical Report, University of Warwick* .

Griffin, J. E. and Brown, P. J. (2010), Inference with normal-gamma prior distributions in regression problems, *Bayesian Analysis* **5**(1), 171–188.

Grünwald, P. D. (2007), *The minimum description length principle*, The MIT Press.

Grünwald, P. D. and de Rooij, S. (2005), Asymptotic log-loss of prequential maximum likelihood codes, *in* 'Proceedings of the Eighteenth Annual Conference on Computational Learning Theory (COLT 2005)', pp. 652–667.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. and Martin, J. B. (1983), A polymorphic DNA marker genetically linked to Huntington's disease, *Nature* **308**, 234–238.

Hannan, E. J. and Quinn, B. G. (1979), The determination of the order of an autoregression, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **41**(2), 190–195.

Havelin, L. I., Espehaug, B., Vollset, S. E. and Engesæter, L. B. (1995), The effect of the type of cement on early revision of Charnley total hip prostheses, *The Journal of Bone & Joint Surgery* **77-A**(10), 1543–1550.

Hayashi, K., Konishi, T. and Kawamoto, T. (2016), A tractable fully Bayesian method for the stochastic block model. arXiv:1602.02256v1.

Hoffman, E. B., Sen, P. K. and Weinberg, C. R. (2001), Within-cluster resampling, *Biometrika* **88**(4), 1121–1134.

Hoffmann, M., Rousseau, J. and Schmidt-Hieber, J. (2015), On adaptive posterior concentration rates, *Ann. Statist.* **43**(5), 2259–2295.

Hofman, J. M. and Wiggins, C. H. (2008), Bayesian approach to network modularity, *Physical Review Letters* **100**, 258701.

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983), Stochastic blockmodels: First steps, *Social Networks* **5**, 109–137.

Holt, J. (1978), Competing risks analyses with special reference to matched pair experiments, *Biometrika* **65**(1), 159–165.

Hooper, G. J., Rothwell, A. G., Hooper, N. M. and Frampton, C. (2012), The relationship between the American Society of Anesthesiologists physical rating and outcome following total hip and knee arthroplasty, *The Journal of Bone & Joint Surgery* **94**(12), 1065–1070.

van Houwelingen, H. C. (2007), Dynamic prediction by landmarking in event history analysis, *Scandinavian Journal of Statistics* **34**, 70–85.

Jiang, W. and Zhang, C.-H. (2009), General maximum likelihood empirical Bayes estimation of normal means, *The Annals of Statistics* **37**(4), 1647–1684.

Jin, J. (2015), Fast community detection by SCORE, *The Annals of Statistics* **43**(1), 57–89.

John, L. K., Loewenstein, G. and Prelec, D. (2012), Measuring the prevalence of questionable research practices with incentives for truth telling, *Psychological Science* **23**(5), 524–532.

Johnsen, S., Sørensen, H., Lucht, U., Søballe, K., Overgaard, S. and Pedersen, A. (2006), Patient-related predictors of implant failure after primary total hip replacement in the initial, short- and long-terms. a nationwide Danish follow-up study including 36 984 patients, *Bone and Joint Journal* **88-B**(10), 1303–1308.

Johnson, V. E. and Rossell, D. (2010), On the use of non-local prior densities in Bayesian hypothesis tests, *Journal of the Royal Statistical Society. Series B (Methodological)* **72**(2), 143–170.

Johnstone, I. M. and Silverman, B. W. (2004), Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences, *The Annals of Statistics* **32**(4), 1594–1649.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The statistical analysis of failure time data. Second edition*, Wiley.

Karrer, B. and Newman, M. E. J. (2011), Stochastic blockmodels and community structure in networks, *Physical Review E* **83**, 016107.

Kass, R. and Raftery, A. E. (1995), Bayes factors, *Journal of the American Statistical Association* **90**(430), 773–795.

Keurentjes, J., Fiocco, M., Schreurs, B., Pijls, B., Nouta, K. and Nelissen, R. (2012), Revision surgery is overestimated in hip replacement., *Bone & Joint Research* **26**, 2389–2430.

Klein, J. P. and Moeschberger, M. L. (2003), *Survival Analysis. Techniques for censored and truncated data. Second edition*, Springer.

Koenker, R. (2014), A Gaussian compound decision bakeoff, *Stat* **3**(1), 12–16.

Koenker, R. and Mizera, I. (2014), Convex optimization, shape constraints, compound decisions and empirical Bayes rules, *Journal of the American Statistical Association* **109**(506), 674–685.

Latouche, A., Allignol, A., Beyersmann, J., Labopin, M. and Fine, J. P. (2013), A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions, *Journal of Clinical Epidemiology* **66**(6), 648 – 653.

Latouche, A., Boisson, V., Chevret, S. and Porcher, R. (2007), Misspecified regression model for the subdistribution hazard of a competing risk, *Statistics in Medicine* **26**(5), 965–974.

Lauritzen, S. (2012), [Catching up faster by switching sooner] : Discussion, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74**(3), 401–402.

Leeb, H. and Pötscher, B. M. (2005), Model selection and inference: Facts and fiction, *Econometric Theory* **21**(1), 21–59.

Lei, J. and Rinaldo, A. (2015), Consistency of spectral clustering in stochastic block models, *The Annals of Statistics* **43**(1), 215–237.

Lévesque, L., Hanley, J., Kezouh, A. and Suissa, S. (2010), Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes, *The BMJ* **340**, b5087.

Lewis, A. S. and Knowles, G. (1992), Image compression using the 2-d wavelet transform, *IEEE Transactions on Image Processing* **1**(2), 244–250.

Lhéritier, A. and Cazals, F. (2015), A sequential nonparametric two-sample test, Technical Report Research Report 8704, INRIA, Sophia Antipolis.

Li, J., Scheike, T. H. and Zhang, M.-J. (2015), Checking Fine and Gray subdistribution hazards model with cumulative sums of residuals, *Lifetime Data Analysis* **21**(2), 197–217.

Li, K.-C. (1989), Honest confidence regions for nonparametric regression, *The Annals of Statistics* **17**(3), 1001–1008.

Lie, S. A., Engesæter, L. B., Havelin, L. I., Gjessing, H. K. and Vollset, S. E. (2004), Dependency issues in survival analyses of 55782 primary hip replacements from 47355 patients, *Statistics in Medicine* **23**, 3227–3240.

Lin, D. (1997), Non-parametric inference for cumulative incidence functions in competing risks studies, *Statistics in Medicine* **16**, 901–910.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993), Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika* **80**(3), 557–572.

Liu, H. and Yu, B. (2013), Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression, *Electronic Journal of Statistics* **7**, 3124–3169.

LROI (2014), Arthroplasty in the picture. Annual report 2014.

Mahomed, N., Barrett, J., Katz, J., Phillips, C., Losina, E., Lew, R., Guadagnoli, E., Harris, W., Poss, R. and Baron, J. (2003), Rates and outcomes of primary and revision total hip replacement in the United States Medicare population, *The Journal of Bone & Joint Surgery* **85**, 27–32.

Makalic, E. and Schmidt, D. F. (2015), A simple sampler for the horseshoe estimator. arXiv:1508.03884.

Makela, K., Matilainen, M., Pulkkinen, P., Fenstad, A., Havelin, L., Engesaeter, L., Furnes, O., Pedersen, A., Overgaard, S., Kärrholm, J., Malchau, H., Garellick, G., Ranstam, J. and Eskelinen, A. (2014), Failure rate of cemented and uncemented total hip replacements: register study of combined Nordic database of four nations, *The BMJ* **348**, f7592.

Mallory, T. H., Lombardi, A., Fada, R., Herrington, S. and Eberle, R. (1999), Dislocation after total hip arthroplasty using the anterolateral abductor split approach, *Clinical Orthopaedics and Related Research* **358**, 166–172.

Martin, R. and Walker, S. G. (2014), Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector, *Electronic Journal of Statistics* **8**(2), 2188–2206.

Maurer, T., Ochsner, P., Schwarzer, G. and Schumacher, M. (2001), Increased loosening of cemented straight stem prostheses made from titanium alloys. an analysis and comparison with prostheses made of cobalt-chromium-nickel alloy, *International Orthopaedics* **25**, 77–80.

McDaid, A. F., Brendan Murphy, T., Friel, N. and Hurley, N. J. (2013), Improved Bayesian inference for the stochastic block model with application to large networks, *Computational Statistics and Data Analysis* **60**, 12–31.

McKeague, I. W., Gilbert, P. B. and Kanki, P. J. (2001), Omnibus tests for comparison of competing risks with adjustment for covariate effects, *Biometrics* **57**(3), 818–828.

Miller, P. D. (2006), *Applied asymptotic analysis*, Vol. 75 of *Graduate Studies in Mathematics*, The American Mathematical Society.

Mitchell, T. J. and Beauchamp, J. J. (1988), Bayesian variable selection in linear regression, *Journal of the American Statistical Association* **83**(404), 1023–1032.

Möllenhoff, G., Walz, M., Muhr, G. and Rehn, J. (1994), Doppelseitige Hüftgelenken endoprothesen: das Zeitintervall als prognostischer Parameter, *Unfallchirurg* **97**, 430–434.

Morris, R. W. (1993), Bilateral procedures in randomised controlled trials, *The Journal of Bone & Joint Surgery* **75-B**, 675–676.

Mossel, E., Neeman, J. and Sly, A. (2012), Reconstruction and estimation in the planted partition model. arXiv:11202.1499v4.

Newman, M. and Girvan, M. (2004), Finding and evaluating community structure in networks, *Physical Review E* **69**, 026113.

Nickl, R. and van de Geer, S. (2013), Confidence sets in sparse regression, *The Annals of Statistics* **41**(6), 2852–2876.

Nickl, R. and Szabó, B. (2014), A sharp adaptive confidence ball for self-similar functions. To appear in Stochastics Processes and their Applications.

NJR (2015), National Joint Registry for England, Wales, Northern Ireland and the Isle of Man. 12th annual report 2015.

Nowicki, K. and Snijders, T. A. B. (2001), Estimation and prediction for stochastic blockstructures, *Journal of the American Statistical Association* **96**(455), 1077–1087.

Ong, K., Mowat, F., Chan, N., Lau, E., Halpern, M. and Kurtz, S. (2006), Economic burden of revision hip and knee arthroplasty in Medicare enrollees, *Clinical Orthopaedics and Related Research* **446**, 22–28.

Pabinger, C. and Geissler, A. (2014), Utilization rates of hip arthroplasty in OECD countries, *Ostheoarthritis and Cartilage* **22**, 734–741.

Pabinger, C., Lothaller, H. and Geissler, A. (2015), Utilization rates of knee-arthroplasty in OECD countries, *Ostheoarthritis and Cartilage* **23**, 1664–1673.

Park, T. and Casella, G. (2008), The Bayesian lasso, *Journal of the American Statistical Association* **103**(482), 681–686.

Park, Y. and Bader, J. S. (2012), How networks change with time, *Bioinformatics* **28**(12), i40–i48.

van der Pas, S., Kleijn, B. and van der Vaart, A. (2014), The horseshoe estimator: Posterior concentration around nearly black vectors, *Electronic Journal of Statistics.* **8**, 2585–2618.

van der Pas, S. L. (2013), Almost the best of three worlds. The switch model selection criterion for single-parameter exponential families, Master's thesis, Leiden University.

Pati, D. and Bhattacharya, A. (2015), Optimal Bayesian estimation in stochastic block models. arXiv:1505.06794.

Pericchi, L. R. and Smith, A. F. M. (1992), Exact and approximate posterior moments for a normal location parameter, *Journal of the Royal Statistical Society. Series B (Methodological)* **54**(3), 793–804.

Picard, D. and Tribouley, K. (2000), Adaptive confidence interval for pointwise curve estimation, *The Annals of Statistics* **28**(1), 298–335.

Polson, N. G. and Scott, J. G. (2010), Shrink globally, act locally: Sparse Bayesian regularization and prediction, *in* J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds, 'Bayesian Statistics 9', Oxford University Press.

Polson, N. G. and Scott, J. G. (2012*a*), Good, great or lucky? Screening for firms with sustained superior performance using heavy-tailed priors, *The Annals of Applied Statistics* **6**(1), 161–185.

Polson, N. G. and Scott, J. G. (2012*b*), On the half-Cauchy prior for a global scale parameter, *Bayesian Analysis* **7**(4), 887–902.

Pratt, J. W. (1962), On the foundations of statistical inference: Discussion, *Journal of the American Statistical Association* pp. 307–326. Discussion.

Prokopetz, J., Losina, E., Bliss, R., Wright, J., Baron, J. and Katz, J. (2012), Risk factors for revision of primary total hip arthroplasty: a systematic review, *BMC Musculosketal Disorders* **13**, 251.

Putter, H., Fiocco, M. and Geskus, R. (2007), Tutorial in biostatistics: competing risks and multi-state models, *Statistics in Medicine* **26**, 2389–2430.

Ramdas, A. and Balsubramani, A. (2015), Sequential nonparametric testign with the law of the iterated logarithm. arXiv:1506.03488.

Ranstam, J., Kärrholm, J., Pulkkinen, P., Keijo, M., Espehaug, B., Pedersen, A. B., Mehnert, F. and Furnes, O. (2011), Statistical analysis of arthroplasty data. II. Guidelines, *Acta Orthopaedica* **82**(3), 258–267.

Ray, K. (2014), Adaptive Bernstein-von Mises theorems in Gaussian white noise, *ArXiv e-prints* .

Ripatti, S. and Palmgren, J. (2000), Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics* **56**(4), 1016–1022.

Robbins, H. (1955), A remark on Stirling's formula, *The American Mathematical Monthly* **62**(1), 26–29.

Robbins, H. (1956), An empirical Bayes approach to statistics, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, volume 1: Contributions to the theory of statistics', University of California Press, Berkeley, California, pp. 157–163.

Robertsson, O. and Ranstam, J. (2003), No bias of ignored bilaterality when analysing the revision risk of knee prostheses: Analysis of a population based sample of 44590 patients with 55298 knee prostheses from the national Swedish Knee Arthroplasty Register, *BMC Musculosketal Disorders* **4**(1).

Robins, J. and van der Vaart, A. (2006), Adaptive nonparametric confidence sets, *The Annals of Statistics* **34**(1), 229–253.

Rohe, K., Chatterjee, S. and Yu, B. (2011), Spectral clustering and the high-dimensional stochastic blockmodel, *The Annals of Statistics* **39**(4), 1878–1915.

Ročková, V. (2015), Bayesian estimation of sparse signals with a continuous spike-and-slab prior. Submitted manuscript, available at http://stat.wharton.upenn.edu/~vrockova/rockova2015.pdf.

Saldana, D. F., Yu, Y. and Feng, Y. (2014), How many communities are there? arXiv:1412.1684v1.

Sanborn, A. N. and Hills, T. T. (2014), The frequentist implications of optional stopping on Bayesian hypothesis tests, *Psychonomic bulletin & review* **21**(2), 283–300.

Sarkar, P. and Bickel, P. J. (2015), Role of normalization in spectral clustering for stochastic blockmodels, *The Annals of Statistics* **43**(3), 962–990.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics* **6**(2), 461–464.

Schwarzer, G., Schumacher, M., Maurer, T. B. and Ochsner, P. E. (2001), Statistical analysis of failure times in total joint replacement, *Journal of Clinical Epidemiology* **54**, 997–1003.

Scott, J. G. (2010), Parameter expansion in local-shrinkage models. arXiv:1010.5265.

Scott, J. G. (2011), Bayesian estimation of intensity surfaces on the sphere via needlet shrinkage and selection, *Bayesian Analysis* **6**(2), 307–328.

Scott, J. G. and Berger, J. O. (2010), Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem, *The Annals of Statistics* **38**(5), 2587–2619.

Serra, P. and Krivobokova, T. (2014), Adaptive empirical Bayesian smoothing splines. arXiv:1411.6860.

Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011), Test martingales, Bayes factors and p-values, *Statistical Science* **26**(1), 84–101.

Shafer, R. E. (1966), Elementary problems. Problem E 1867, *The American Mathematical Monthly* **73**(3), 309.

Shao, J. (1997), An asymptotic theory for linear model selection, *Statistica Sinica* **7**, 221–264.

SHAR (2014), The Swedish Hip Arthroplasty Register. Annual report 2014.

Silver, M., Janousova, E., Hua, X., Thompson, P. M. and Montana, G. (2012), Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression, *NeuroImage* **63**(3), 1681 – 1694.

Sniekers, S. and van der Vaart, A. (2015*a*), Adaptive Bayesian credible sets in regression with a Gaussian process prior, *Electronic Journal of Statistics* **9**(2), 2475–2527.

Sniekers, S. and van der Vaart, A. (2015*b*), Adaptive credible bands in nonparametric regression with Brownian motion prior. arXiv:1504.07972.

Sniekers, S. and van der Vaart, A. (2015*c*), Credible sets in the fixed design model with Brownian motion prior, *Journal of Statistical Planning and Inference* **166**, 78–86.

Snijders, T. A. and Nowicki, K. (1997), Estimation and prediction for stochastic blockmodels for graphs with latent block structure, *Journal of Classification* **14**, 75–100.

Stone, M. (1977), An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **39**(1), 44–47.

Suissa, S. (2007), Immortal time bias in observational studies of drug effects, *Pharmacoepidemiology and Drug Safety* **16**(3), 241–249.

Suwan, S., Lee, D. S., Tang, R., Sussman, D. L., Tang, M. and Priebe, C. E. (2016), Empirical Bayes estimation for the stochastic blockmodel, *Electronic Journal of Statistics* **10**, 761–782.

Sylvestre, M., Huszti, E. and Hanley, J. (2006), Do Oscar winners live longer than less successful peers? A reanalysis of the evidence., *Annals of Internal Medicine* **145**(5), 361–363.

Szabó, B., van der Vaart, A. W. and van Zanten, J. H. (2015*a*), Frequentist coverage of adaptive nonparametric Bayesian credible sets, *Ann. Statist.* **43**(4), 1391–1428.

Szabó, B., van der Vaart, A. and van Zanten, H. (2015*b*), Honest Bayesian confidence sets for the L2-norm, *Journal of Statistical Planning and Inference* **166**, 36 – 51. Special Issue on Bayesian Nonparametrics.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tsui, L.-C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R., Schumm, J. W., Eiberg, Hans en Mohr, J., Kennedy, D., Plavsic, N., Zsiga, M., Markiewicz, D., Akots, G., Brown, V., Helms, C., Gravius, T., Parker, C., Rediker, K. and Donis-Keller, H. (1985), Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker, *Science* **230**, 1054–1057.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

van der Vaart, A. and van Zanten, H. (2009), Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth, *The Annals of Statistics* **37**(5B), 2655–2675.

van der Pas, S., Scott, J., Chakraborty, A. and Bhattacharya, A. (2016), *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.1.0.

van der Vaart, A. W. (1998), *Asymptotic statistics*, Cambridge University Press.

Visuri, T., Turula, K. B., Pulkkinen, P. and Nevalainen, J. (2002), Survivorship of hip prosthesis in primary arthrosis. Influence of bilaterality and interoperative time in 45000 hip prostheses from the Finnish Endoprosthesis Register, *Acta Orthopaedica Scandinavina* **73**(3), 287–290.

Wagenmakers, E.-J. (2007), A practical solution to the pervasive problems of p-values, *Psychonomic Bulletin & Review* **14**(5), 779–804.

Wang, Y. X. R. and Bickel, P. J. (2015), Likelihood-based model selection for stochastic block models. arXiv:1502.02069v1.

Wienke, A. (2003), Frailty models, Technical Report WP-2003-032, Max Planck Institute for Demographic Research.

Wilcock, G. (1978), Benefits of total hip replacement to older patients and the community, *British Medical Journal* **2**, 37–39.

Yang, Y. (2005), Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation, *Biometrika* **92**(2), 937–950.

Yang, Y., Wainwright, M. J. and Jordan, M. I. (2015), On the computational complexity of high-dimensional Bayesian variable selection. arXiv:1505.07925.

Yuan, M. and Lin, Y. (2005), Efficient empirical Bayes variable selection and estimation in linear models, *Journal of the American Statistical Association* **100**(472), 1215–1225.

Zachary, W. W. (1977), An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**(4), 452–473.

Zhang, A. Y. and Zhou, H. H. (2015), Minimax rates of community detection in stochastic block models. Preprint available at http://www.stat.yale.edu/~hz68/CommunityDetection.pdf.

Zhang, C.-H. and Zhang, S. S. (2014), Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**(1), 217–242.

Zhao, Y., Levina, E. and Zhu, J. (2012), Consistency of community detection in networks under degree-corrected stochastic block models, *The Annals of Statistics* **40**(4), 2266–2292.

Zhou, B., Fine, J., Latouche, A. and Labopin, M. (2012), Competing risks regression for clustered data, *Biostatistics* **13**(3), 371–383.

Samenvatting

De eerste drie hoofdstukken van dit proefschrift betreffen problemen waarin er maar weinig signalen zijn, temidden van heel veel ruis. Te denken valt aan het vinden van genen geassocieerd met een bepaalde ziekte, het detecteren van supernovae in astronomische opnamen, of het comprimeren van afbeeldingen. In het bijzonder wordt in dit proefschrift aandacht besteed aan een simpel model, waarin elke waarneming gelijk is aan de som van een parameter, en normaal verdeelde ruis. Aangenomen wordt dat de meeste van die parameters gelijk aan nul zijn. Het doel is tweeledig: het schatten van de parameters, en de onzekerheid in die schatting aangeven.

Dit probleem wordt aangepakt met Bayesiaanse statistiek. Daartoe wordt een *a-priori*-verdeling op de parameters aangenomen. Na waarneming van de observaties leidt dit tot een *a-posteriori*-verdeling op de parameters, welke gebruikt wordt om tot een schatting van de parameters te komen. In dit proefschrift wordt de *hoefijzer-a-priori*-verdeling (Carvalho et al., 2010) bestudeerd. De verwachtingswaarde van de bijbehorende *a-posteriori*-verdeling wordt als schatter van de parameters gebruikt, en noemen we de hoefijzerschatter. De hoefijzerverdeling hangt af van een parameter τ. Hoe kleiner τ gekozen wordt, hoe dichter de uiteindelijke schattingen bij nul komen te liggen.

Het belangrijkste resultaat uit Hoofdstuk 1 is dat de hoefijzerschatter geschikt is om de parameters mee te schatten (in de zin van minimax-optimaliteit) indien τ klein genoeg wordt gekozen. Hoe groot τ precies moet zijn, hangt af van het aantal signalen onder de parameters. De resultaten uit Hoofdstuk 1 leidden tot de vraag of de goede eigenschappen voorbehouden zijn aan de hoefijzerverdeling. In Hoofdstuk 2 wordt aangetoond dat dat niet het geval is: de resultaten kunnen herhaald worden voor een hele klasse aan verdelingen.

De aandacht in de eerste twee hoofdstukken gaat voornamelijk uit naar het schatten van de parameters. In Hoofdstuk 3 wordt ingegaan op de kwestie van de onzekerheid in deze schattingen. Aan de hand van de hoefijzer-*a-posteriori*-verdeling wordt er een bereik aan waarden voor de parameters vastgesteld, en gekeken hoe vaak de echte parameterwaarde in dat bereik ligt, en of dat bereik niet onnodig groot is. In Hoofdstuk 3 wordt bewezen dat de hoefijzer *a-posteriori*-verdeling geschikt is om onzekerheid in de parameters uit te drukken, tenzij de echte waarde van de parameter dichtbij de "universele drempelwaarde" van $\sqrt{2 \log n}$ ligt, waarbij n het aantal waarnemingen is. Voor deze gunstige eigenschappen is het van belang dat de parameter τ goed gekozen wordt. Dit blijkt te kunnen door ofwel een *a-priori*-verdeling op τ te plaatsen, of door een schatter te gebruiken die in Hoofdstuk 3 bestudeerd wordt. In beide gevallen is het niet nodig om op

voorhand te weten hoeveel signalen er zijn.

Hoofdstuk 4 betreft een ander probleem, namelijk het vinden van groepen in een netwerk dat is ontstaan volgens het *stochastisch blokmodel*. Als we een sociaal netwerk als voorbeeld nemen, zouden we observeren wie met wie bevriend is, en aannemen dat de kans op een vriendschap tussen twee personen alleen afhangt van de groep waar ieder lid van is. Het doel is om te achterhalen wie lid is van welke groep. In Hoofdstuk 4 wordt hiervoor wederom een Bayesiaanse aanpak gebruikt. De kans dat de bestudeerde schatter de groepen correct identificeert gaat naar één wanneer het aantal individuen toeneemt, mits het aantal connecties tussen individuen niet te klein is.

In Hoofdstuk 5 wordt het *switch-criterium*, een nieuwe methode om een hypothesetoets uit te voeren, geëvalueerd op drie eigenschappen:

1. Wordt de juiste hypothese gekozen?
2. Hoe goed worden de parameters behorend bij de hypotheses geschat?
3. Is het criterium gevoelig voor de stopregel?

Dit laatste is een probleem bij de meeste klassieke hypothesetoetsen. Wanneer maar lang genoeg wordt doorgegaan met waarnemingen doen, zal de nulhypothese uiteindelijk altijd verworpen worden, ongeacht of deze waar is of niet. In een enquête onder psychologen gaf 55% van de deelnemers toe wel eens pas te besluiten of er meer waarnemingen gedaan zouden worden na het zien van de eerste resultaten (John et al., 2012). Het zou beter bij de wetenschappelijke praktijk passen, als dergelijk gedrag geen problemen zou opleveren voor de validiteit van de statistische analyse. Met het switch-criterium is dat het geval, mits de nulhypothese een punthypothese is, en er daarvoor dus geen verdere parameters geschat hoeven te worden. Dit gaat ten koste van de precisie bij het schatten van parameters behorend bij de overige hypotheses. De eerste eigenschap, consistentie, komt niet in het geding. Dit geldt voor hypotheses die in elkaar bevat zitten, wanneer de gepostuleerde verdelingen een exponentiële familie vormen.

Hoofdstuk 6 is toegepast van aard. Het is gebaseerd op analyse van data over heupprothesen, afkomstig van het LROI (Landelijke Registratie Orthopedische Implantaten). De vraag is hoe lang het duurt tot er een nieuwe ingreep aan de heupprothese (revisie) plaatsvindt, en welke eigenschappen van de patiënt (zoals leeftijd en diagnose) daarmee geassocieerd zijn. De statistische analyse wordt bemoeilijkt door drie problemen.

De eerste twee komen door de aanwezigheid van *bilaterale patiënten*, die aan beide kanten een heupprothese hebben. Twee waarnemingen uit één patiënt zijn afhankelijk, terwijl voor de meeste methodes onafhankelijkheid wordt aangenomen. De tweede moeilijkheid is dat er doorgaans enkele maanden of jaren tussen het plaatsen van de twee heupprotheses zit. Wanneer deze tijdsafhankelijkheid niet goed wordt meegenomen, kan dat onbedoelde effecten op de uitkomsten van de analyse hebben. Zulke effecten treden ook op bij de bewering dat Oscar-winnaars langer leven, waarbij over het hoofd wordt gezien dat iemand in leven moet zijn om een Oscar te krijgen, en dus een minimum aantal jaren moet overleven, terwijl die overlevingseis niet geldt voor de mensen waarmee vergeleken wordt (Sylvestre et al., 2006). De derde moeilijkheid is dat een patiënt kan overlijden voordat revisie kan plaatsvinden. Het risico hierop is aanzienlijk, aangezien patiënten gemiddeld ongeveer 69 jaar oud zijn wanneer ze een heupprothese krijgen (LROI, 2014).

Deze drie complicaties worden besproken in Hoofdstuk 6, waarna enkele voorlopige resultaten op de data van de LROI worden gepresenteerd.

Dankwoord

Possunt, quia posse videntur.
Zij kunnen het, omdat ze denken het te kunnen.
– Vergilius, *Aeneis* 5.231

Met veel plezier grijp ik deze gelegenheid aan om mijn dank te betuigen aan iedereen die mij heeft aangemoedigd tijdens de totstandbrenging van dit proefschrift.

Aad, ik heb heel veel respect voor je, en er is zoveel om je voor te bedanken. Dank voor de inspirerende samenwerking waarbij je nooit genoegen nam met een resultaat tot helder was waarom het waar was, en of het nog beter kon. Dank voor je mentorschap, en het me gedecideerd tegenspreken wanneer ik beweerde iets niet te kunnen. Dank voor de vele mogelijkheden om congressen te bezoeken, en daar nieuwe contacten te leggen. En boven alles, dank voor de complete vrijheid die je me de afgelopen jaren hebt gegeven om mijn eigen interesses te volgen.

Deze interesses had ik nooit na kunnen jagen zonder mijn co-auteurs. Daarvoor wil ik hen graag bedanken. Bas, dankzij jouw enthousiaste doch realistische betoog tijdens een lunch vier jaar geleden heb ik de knoop doorgehakt en besloten deze uitdaging aan te gaan. Daar ben ik je zeer erkentelijk voor. Peter, jou wil ik graag bedanken voor je gedegen begeleiding van mijn eerste stappen in het wiskundig onderzoek. Je creatieve ideeën en optimisme gaven telkens aanleiding om weer verder te gaan. Marta, ik had me geen betere introductie in de biostatistiek kunnen wensen, en ik hoop dat dit slechts het begin is van een langdurige samenwerking. Rob, dank voor de geduldige en enthousiaste uitleg over totale heupprotheses, en de mogelijkheid om deel te nemen aan het ISAR congres. Johannes, dank voor de samenwerking. Je intensieve manier van samenwerken vond ik heel prettig, en leidde snel tot mooie resultaten. Ik heb veel van je geleerd. Dank ook voor de vele kopjes thee, en de soeprecensies. J-B, it was a pleasure to have the opportunity to work with you. I have fond memories of our hours spent doing computations on a blackboard. Botond, thank you for the productive collaboration, and your enthusiasm throughout. Your unwillingness to give up is truly something else. Richard, dank voor de vele interessante gesprekken en het delen van je expertise.

Verder wil ik graag mijn collega's bedanken voor de leuke tijd op het MI. Suzanne, jij had het pad al een heel stuk bewandeld toen ik begon. Wat ben ik blij dat je me vanaf het begin vergezeld hebt naar congressen, en er altijd was om lief, leed, paniekmomenten en rare ervaringen met Mexicaanse artsen mee te delen. Fengnan, thank you for sharing all kinds of advice with me and never giving up on finding a type of Chinese tea that I would

like. My time at the MI would not have been the same without you. Maarten, Sanne en Vincent, ik heb het erg getroffen met jullie als kantoorgenoten. De gezamenlijke strijd tegen het *panic monster* en de *instant gratification monkey* was niet altijd even effectief, maar wel heel gezellig. Tim, dank voor de leerzame gesprekken, het delen van je ervaringen, en eindeloze begrip wanneer ik in de problemen raakte met de spreekwoordelijke hooi en vork.

Dank ook aan alle andere MI'ers voor het goede gezelschap in de *salon du thé*, common room, en aan de lunchtafel. Björn, Lotte, Stefanie, Frejanne, Frank, Maja, Willem, Laurens, Kolyan, Anja, Abtien, Gino, Kevin, Dino, Dong, Giulia, Frits, Eric, Shota, Corine, Carlo - bedankt voor alle gezellige en gedenkwaardige momenten.

De steun van de niet-wiskundigen in mijn leven kan niet overschat worden. Ik wil graag mijn vrienden en familie bedanken voor hun onwrikbare overtuiging "dat het wel goed zou komen met dat proefschrift". Ik ben dankbaar en ontroerd dat iedereen zo in mij gelooft en enthousiast al mijn plannen aanmoedigt, hoe veranderlijk en wisselvallig die soms ook mogen zijn.

Leiden, januari 2017

Curriculum Vitae

De auteur van dit proefschrift is geboren op 10 maart 1989 te Hilversum. In 2006 behaalde ze het vwo-diploma aan het Gemeentelijk Gymnasium Hilversum. In datzelfde jaar begon ze aan de studie Geneeskunde aan de Universiteit van Amsterdam, die ze na het behalen van de propedeuse (*cum laude*) alweer beëindigde. In 2007 begon ze aan de Universiteit Leiden aan de studies Wiskunde en Griekse en Latijnse Taal en Cultuur, die ze respectievelijk in 2013 en 2012 afrondde, beide *cum laude*. Tijdens de masters bracht ze een semester door aan The University of British Columbia in Vancouver. De masterscriptie die voortkwam uit de studie Wiskunde, getiteld *Almost the best of three worlds: the switch model selection criterion for single-parameter exponential families* en geschreven onder begeleiding van prof. dr. Peter Grünwald, werd bekroond met de ASML Afstudeerprijs voor Wiskunde van de Koninklijke Hollandse Maatschappij der Wetenschappen en met de Leidse Universitaire Scriptieprijs.

Aansluitend aan haar afstuderen is ze begonnen aan het onderzoek dat heeft geleid tot dit proefschrift, onder begeleiding van prof. dr. Aad van der Vaart, aan het Mathematisch Instituut te Leiden. Gedurende die periode organiseerde ze het PhD colloquium voor promovendi van het MI en was ze een jaar voorzitter van de Nederlandse sectie van European Women in Mathematics.