Author: Pas, S.L. van der
Title: Topics in mathematical and applied statistics
Issue Date: 2017-02-28

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

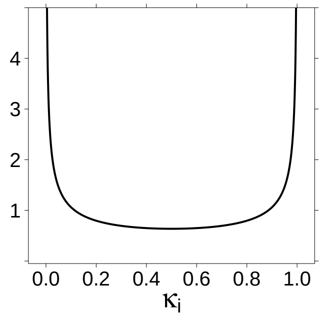

Figure 1: Prior density of κ_i for $\tau = 1$.

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 $i = 1, \dots, n$, 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.