Chapter 3

Bayesian Inconsistency under Misspecification

We empirically show that Bayesian inference can be inconsistent under misspecification in simple linear regression problems, both in a model averaging/selection and in a Bayesian ridge regression setting. We use the standard linear model, which assumes homoskedasticity, whereas the data are heteroskedastic (though significantly, there are no outliers), and observe that the posterior puts its mass on ever more high-dimensional models as the sample size increases. To remedy the problem, we equip the likelihood in Bayes' theorem with an exponent called the learning rate, and we propose the *SafeBayesian* method to learn the learning rate from the data. SafeBayes tends to select small learning rates as soon the standard posterior is not 'cumulatively concentrated', and its results on our data are quite encouraging.

In this chapter, we focus on introducing both the problem and the solution we propose, and we provide our main experiments with Bayes and SafeBayes. The discussion of Bayesian inconsistency will be continued in Chapters 4 (analysing the underlying reasons for the behaviour of Bayes and SafeBayes) and 5 (providing several additional experiments to check the generality of our findings). An overview of these three chapters is provided in Section 3.1.1, and an 'executive summary' of all the experiments in Chapters 3 and 5 combined is provided in Section 3.5.5 at the end of Chapter 3.

3.1 Introduction

The problem We empirically demonstrate the inconsistency of Bayes factor model selection, model averaging and Bayesian ridge regression under model misspecification on a simple linear regression problem with random design. We sample data $(X_1, Y_1), (X_2, Y_2), \ldots$ i.i.d. from a distribution P^*, where $X_i = (X_{i1}, \ldots, X_{ip_{\max}})$ are high-dimensional vectors, and we allow $p_{\max} = \infty$. We use nested models $\mathcal{M}_0, \mathcal{M}_1, \ldots$ where $\mathcal{M}_p$ is a standard linear model, consist-

ing of conditional distributions $P(\cdot \mid \beta, \sigma^2)$ expressing that

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i \tag{3.1}$$

is a linear function of $p \leq p_{\max}$ covariates with additive independent Gaussian noise $\epsilon_i \sim N(0, \sigma^2)$. We equip each of these models with standard priors on coefficients and the variance, and also put a discrete prior on the models themselves. $\mathcal{M} := \bigcup_{p=0,\ldots,p_{\max}} \mathcal{M}_p$ does not contain the conditional 'ground truth' $P^*(Y \mid X)$ (hence the model is 'misspecified'), but it does contain a $\tilde{P}$ that is 'best' in several respects: it is closest to P^* in KL (Kullback-Leibler) divergence, it represents the true regression function (leading to the best squared error loss predictions among all $P \in \mathcal{M}$) and it has the true marginal variance (explained in Section 3.2.3). Yet, while $\tilde{P} \in \mathcal{M}_0$ and $\mathcal{M}_0$ receives substantial prior mass, as n increases, the posterior puts most of its mass on complex $\mathcal{M}_p$'s with higher and higher p's, and, conditional on these $\mathcal{M}_p$'s, at distributions which are very far from P^* both in terms of KL divergence and in terms of L_2 risk, leading to bad predictive behaviour in terms of squared error. Figures 3.1 and 3.2 illustrate a particular instantiation of our results, obtained when X_{ij} are polynomial functions of S_i and $S_i \in [-1, 1]$ uniformly i.i.d. We also show comparably bad predictive behaviour for various versions of Bayesian ridge regression, involving just a single, high-but-finite dimensional model. In that case Bayes eventually recovers and concentrates on $\tilde{P}$, but only at a sample size that is incomparably larger than what can be expected if the model is correct.

These findings contradict the folk wisdom that, if the model is incorrect, then "Bayes tends to concentrate on neighbourhoods of the distribution(s) in $\mathcal{M}$ that is/are closest to P^* in KL divergence." Indeed, the strongest actual theorems to this end that we know of, (Kleijn and Van der Vaart, 2006; De Blasi and Walker, 2013; Ramamoorthi et al., 2013), hold, as the authors emphasize, under regularity conditions that are substantially stronger than those needed for consistency when the model is correct (as by e.g. Ghosal et al. (2000) or Zhang (2006a)), and our example shows that consistency may fail to hold even in relatively simple problems.

The solution: Generalized posterior and SafeBayes Bayesian updating can be enhanced with a *learning rate* η, an idea put forward independently by several authors (Vovk, 1990; McAllester, 2003; Barron and Cover, 1991; Walker and Hjort, 2002; Zhang, 2006a) and suggested as a tool for dealing with misspecification by Grünwald (2011; 2012). η trades off the relative weight of the prior and the likelihood in determining the *η-generalized posterior*, where $\eta = 1$ corresponds to standard Bayes and $\eta = 0$ means that the posterior always remains equal to the prior. When choosing the 'right' η, which in our case is significantly smaller than 1 but of course not 0, η-generalized Bayes becomes competitive again. In general, the optimal η depends on the underlying ground truth P^*, and the problem has always been how to determine the optimal η empirically, from the data.

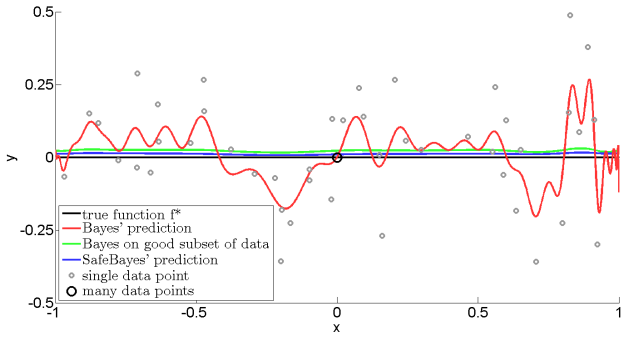

Figure 3.1: The conditional expectation $\mathbf{E}[Y \mid X]$ according to the full Bayesian posterior based on a prior on models $\mathcal{M}_0, \ldots, \mathcal{M}_{50}$ with polynomial basis functions, given 100 data points sampled i.i.d. $\sim P^*$ (about 50 of which are at $(0,0)$). Standard Bayes overfits, not as dramatically as maximum likelihood or unpenalized least squares, but still enough to show dismal predictive behaviour as in Figure 3.2. In contrast, SafeBayes (which chooses learning rate $\eta \approx 0.4$ here) and standard Bayes trained only at the points for which the model is correct (not $(0,0)$) both perform very well.

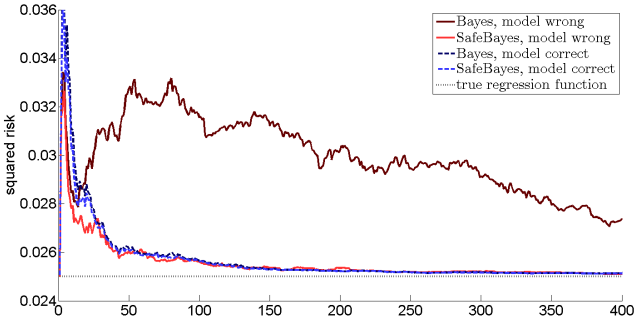

Figure 3.2: The expected squared error risk (defined in (3.3)), obtained when predicting by the full Bayesian posterior (brown curve), the SafeBayesian posterior (red curve) and the optimal predictions (black dotted curve), as a function of sample size for the setting of Figure 3.1. SafeBayes is the R-log-version of SafeBayes defined in Section 3.4.2. Precise definitions and further explanation in Section 3.5.1 and Section 3.5.2.

Recently, Grünwald (2012) proposed the *SafeBayesian* algorithm for learning η, and theoretically showed that it achieves good convergence rates in terms of KL divergence on a variety of problems. Here we show empirically that SafeBayes performs excellently in our regression setting, being competitive with standard Bayes if the model is correct and very significantly outperforming not just standard Bayes, but also cross-validation and approaches such as AIC when the model is incorrect. We do this by providing a wide range of experiments, varying parameters of the problem such as the priors and the true regression function and studying various performance indicators such as the squared error risk, the posterior on the variance etc.

We note that a Bayesian's (and our) first instinct would be to learn η itself in a Bayesian manner instead. Yet this does not solve the problem, as we show in Section 3.5.4, where we consider a setting in which $1/\eta$ turns out to be exactly equivalent to the λ regularization parameter in the Bayesian Lasso and ridge regression approaches. We find that selecting η by (empirical) Bayes, as suggested by e.g. Park and Casella (2008), does not nearly regularize enough in our misspecification experiments. In the Bayesian ridge regression setting with fixed variance, the SafeBayesian algorithm becomes very similar to learning λ by cross-validation with squared-error loss, as is standard in frequentist ridge regression (cross-validation with a logarithmic score does *not* work however). In the varying variance case, there is no such straightforward interpretation of SafeBayes.

The type of misspecification The models are misspecified in that they make the standard assumption of homoskedasticity — σ^2 is independent of X — whereas in reality, under P^*, there is heteroskedasticity, there being a region of X with low and a region with (relatively) high variance. Specifically, in our simplest experiment the 'true' P^* is defined as follows: at each i, toss a fair coin. If the coin lands heads, then sample X_i from a uniform distribution on $[-1, 1]$, and set $Y_i = 0 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_0^2)$. If the coin lands tails, then set $(X_i, Y_i) = (0, 0)$, so that there is no variance at all. The 'best' conditional density $\tilde{P}$, closest to $P^*(Y \mid X)$ in KL divergence, representing the true regression function $Y = 0$ and reliable in the sense of Section 3.2.3, is then given by (3.1) with all β's set to 0 and $\tilde{\sigma}^2 = \sigma_0^2/2$. In a typical sample of length n, we will thus have approximately $n/2$ points with X_i uniform and Y_i normal with mean 0, and approximately $n/2$ points with $(X_i, Y_i) = (0, 0)$. These points seem 'easy' since they lie exactly on the regression function one would hope to learn; but they really wreak severe havoc.

The in-liers cause the problem While it is well-known that in the presence of outliers, Gaussian assumptions on the noise lead to problems, both for frequentist and Bayesian procedures, in the present problem we have *'in-liers'* rather than outliers. Also, if we slightly modify the setup so that homoskedasticity holds, standard Bayes starts behaving excellently, as again depicted in Figures 3.1 and 3.2. Finally, while the figure shows what happens for polynomials, we used independent multivariate X's rather than nonlinear basis functions in the main experiments below, getting essentially the same results. All

this indicates that the inconsistency is really caused by misspecification, in particular the presence of in-liers, and not by anything else. The setup is inspired by the work of Grünwald and Langford (2004, 2007), who gave a mathematical proof that Bayesian inference can be inconsistent under misspecification in a related but much more artificial classification setting. Here we show that this can also happen in a much more natural regression setting. The setting being more natural, it is also harder to analyse, and we only demonstrate the inconsistency empirically.

3.1.1 Overview of Chapters 3 to 5

KL-associated inference tasks Section 3.2 introduces our regression setting and the main concepts needed to understand our results. A crucial point here is that, if Bayesian (or other likelihood-based methods) converge at all to a distribution in the model $\mathcal{M}$, this distribution (often called the 'pseudo-truth') is the $\tilde{P} \in \mathcal{M}$ that minimizes KL divergence to the true distribution P^*. While the minimum KL divergence point is often not of intrinsic interest, for some (not all) models, $\tilde{P}$ can be of interest for other reasons as well (Royall and Tsou, 2003): there may be *associated* inference tasks for which $\tilde{P}$ is suitable as well. For standard linear models with fixed σ^2, the main associated task is squared error prediction: the KL-optimal $\tilde{P}$ is also optimal, among all $P \in \mathcal{M}$, in terms of squared error prediction risk. If additionally σ^2 becomes a free parameter, then it is also reliable, which roughly means that it is optimal in determining its own squared error prediction quality (Section 3.2.3; we have a lot more to say about associated inference tasks in Section 4.3). Thus, whenever one is prepared to work with linear models and one is interested in squared error risk or reliability, then Bayesian inference would seem the way to go, even if one suspects misspecification... at least if there is consistency.

The SafeBayesian algorithm Section 3.3 introduces the η-generalized posterior and instantiates it to the linear model. Section 3.4 introduces the 'SafeBayesian' algorithm, which learns η from the data. This is done via Dawid's (1984) *prequential* view on Bayesian inference. We then provide four instantiations of the SafeBayes method to linear models.

Section 3.5 discusses our experiments. We first provide the necessary preparation in Section 3.5.1 and 3.5.2. Section 3.5.3 gives the results of our first experiment, a comparison of Bayesian and SafeBayesian model averaging and selection in two settings, one with a correct model and one with a model corrupted by 50% easy points as above, but with independent Gaussian rather than polynomial inputs. Section 3.5.4 repeats these experiments for a Bayesian ridge regression setting, Section 3.5.5 provides an 'executive summary'. In all experiments SafeBayesian methods behave much better in terms of squared error risk and reliability than standard Bayes if the model is incorrect, and hardly worse (sometimes still better) than standard Bayes if the model is correct.

Good vs. bad misspecification: Nonconcentration and hypercompression In and of itself, the fact that one obtains inconsistency with homoskedastic

models and heteroskedastic data may not be very surprising; indeed, whether similar phenomena occur in real-world data needs further study. The main strength of our example is rather that it clearly shows what can happen in principle, and indicates how one may go about solving it. We explain this in Section 4.1 in Chapter 4, in particular on the basis of Figure 4.1 on page 74, *the essential picture to understand the phenomenon*. Inconsistency can only arise under a 'bad' form of misspecification, depicted by the figure. Under bad misspecification, the posterior may *fail to concentrate*, and this causes trouble. As a theoretical contribution of this chapter, we show in this section that, under some conditions, a Bayesian strongly believes that her posterior will, in some sense, concentrate fast. Indeed, SafeBayes will only select $\eta \ll 1$ if the standard posterior is nonconcentrated, and may thus be (loosely) viewed as a particular 'prior predictive check'.

Posterior nonconcentration in turn can lead to 'hypercompression', the phenomenon that the Bayes predictive distribution behaves *substantially better* under a logarithmic scoring rule than the best distribution $\tilde{P} \in \mathcal{M}$; this can happen because the Bayes predictive distribution — a mixture of elements of $\mathcal{M}$ — behaves substantially differently from any of the elements of $\mathcal{M}$. Somewhat paradoxically (Section 4.1.3), Bayes' overly good log-loss behaviour is exactly what causes it to perform badly for the associated inference tasks (squared error prediction and reliability, in our case). Thus, there can be an inherent tension between behaviour under log-loss and behaviour under its associated tasks, a discrepancy which one can measure by the *mixability gap* (Section 4.1.4), a theoretical concept introduced by Van Erven et al. (2011) and Grünwald (2012). If one is interested in log-loss, standard Bayes is just fine; the SafeBayesian algorithm should be used if one wants to optimize behaviour against the associated tasks. Of course, whether such a task-dependent modification of Bayes is desirable needs discussion, which we provide in Section 4.3.

Additional experiments In Chapter 5 we provide a battery of experiments to check the robustness of our results. Specifically, we investigate what happens if we vary our models and priors (using e.g. a fixed σ^2 and standard priors used in the regression literature), our methods, and if we vary the data-generating distribution using e.g. 'easy' points that are close to, but not exactly $(0, 0)$. Our main conclusion here is that, of the four versions of SafeBayes which we propose, one is uncompetitive and among the other three, there is no clear winner — although they consistently outperform Bayes under misspecification. Furthermore we show that AIC, BIC and cross-validation also have serious problems in our regression setup.

3.2 Preliminaries

3.2.1 Setting, logarithmic risk, optimal distribution

In this chapter we consider data $Z^n = Z_1, Z_2, \ldots, Z_n \sim$ i.i.d. P^*, where each $Z_i = (X_i, Y_i)$ is an independently sampled copy of $Z = (X, Y)$, X taking val-

ues in some set $\mathcal{X}$, Y taking values in $\mathcal{Y}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We are given a *model* $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$ parameterized by (possibly infinite-dimensional) Θ, and consisting of conditional distributions $P_\theta(Y \mid X)$, extended to n outcomes by independence. For simplicity we assume that all P_θ have corresponding conditional densities f_θ, and similarly, the conditional distribution $P^*(Y \mid X)$ has a conditional f^*, all with respect to the same underlying measure. While we do not assume $P^*(Y \mid X)$ to be in (or even 'close' to) $\mathcal{M}$, we want to learn, from given data Z^n, a 'best' (in a sense to be defined below) element of $\mathcal{M}$, or at least, a distribution on elements of $\mathcal{M}$ that can be used to make predictions about future data. While our experiments focus on linear regression, the discussion in this section holds for general conditional density models. The logarithmic score, henceforth abbreviated to *log-loss*, is defined in the standard manner: the loss incurred when predicting Y based on density $f(\cdot \mid x)$ and Y takes on value y, is given by $-\log f(y \mid x)$. A central quantity in our setup is then the *expected log-loss* or *log-risk*, defined as

$$\textsc{risk}^{\log}(\theta) := \mathbf{E}_{(X,Y)\sim P^*}[-\log f_\theta(Y \mid X)],$$

where here as in the remainder of this chapter, log denotes the natural logarithm.

We let P^*_X be the marginal distribution of X under P^*. The *Kullback-Leibler (KL) divergence* $D(P^* \| P_\theta)$ between P^* and conditional distribution P_θ is defined as the expectation, under $X \sim P^*_X$, of the KL divergence between P_θ and the 'true' conditional $P^*(Y \mid X)$: $D(P^* \| P_\theta) = \mathbf{E}_{X\sim P^*_X}[D(P^*(\cdot \mid X) \,\|\, P_\theta(\cdot \mid X))]$. A simple calculation shows that for any θ, θ',

$$D(P^* \| P_\theta) - D(P^* \| P_{\theta'}) = \textsc{risk}^{\log}(\theta) - \textsc{risk}^{\log}(\theta'),$$

so that the closer P_θ is to P^* in terms of KL divergence, the smaller its log-risk, and the better it is, on average, when used for predicting under the log-loss.

Now suppose that $\mathcal{M}$ contains a unique distribution that is closest, among all $P \in \mathcal{M}$ to P^* in terms of KL divergence. We denote such a distribution, if it exists, by $\tilde{P}$. Then $\tilde{P} = P_\theta$ for at least one $\theta \in \Theta$; we pick any such θ and denote it by $\tilde{\theta}$, i.e. $\tilde{P} = P_{\tilde{\theta}}$, and note that it also minimizes the log-risk:

$$\textsc{risk}^{\log}(\tilde{\theta}) = \min_{\theta\in\Theta} \textsc{risk}^{\log}(\theta) = \min_{\theta\in\Theta} \mathbf{E}_{(X,Y)\sim P^*}[-\log f_\theta(Y \mid X)]. \tag{3.2}$$

We shall call such a $\tilde{\theta}$ *(KL-)optimal*.

Since, in regions of about equal prior density, the log Bayesian posterior density is proportional to the log likelihood ratio, we hope that, given enough data, with high P^*-probability, the posterior puts most mass on distributions that are close to $P_{\tilde{\theta}}$ in KL divergence, i.e. that have log-risk close to optimal. Indeed, all existing consistency theorems for Bayesian inference under misspecification express concentration of the posterior around $P_{\tilde{\theta}}$.

3.2.2 A special case: The linear model

Fix some $p_{\max} \in \{0, 1, \ldots\} \cup \{\infty\}$. We observe data $Z_1, \ldots, Z_n$ where $Z_i = (X_i, Y_i)$, $Y_i \in \mathbf{R}$ and $X_i = (1, X_{i1}, \ldots, X_{ip_{\max}}) \in \mathbf{R}^{p_{\max}+1}$. Note that this is as in (3.1) but from now on we adopt the standard convention to take $X_{0i} \equiv 1$ as a dummy random variable. We denote by $\mathcal{M}_p = \{P_{p,\beta,\sigma^2} \mid (p, \beta, \sigma^2) \in \Theta_p\}$ the standard linear model with parameter space $\Theta_p := \{(p, \beta, \sigma^2) \mid \beta = (\beta_0, \ldots, \beta_p)^\top \in \mathbf{R}^{p+1}, \sigma^2 > 0\}$, where the entry p in (p, β, σ^2) is redundant but included for notational convenience. We let $\Theta = \bigcup_{p=0,\ldots,p_{\max}} \Theta_p$. $\mathcal{M}_p$ states that for all i, (3.1) holds, where $\epsilon_1, \epsilon_2, \ldots \sim$ i.i.d. $N(0, \sigma^2)$. When working with linear models $\mathcal{M}_p$, we are usually interested in finding parameters β that predict well in terms of the *squared error loss function* (henceforth abbreviated to *square-loss*): the square-loss on data (X_i, Y_i) is $(Y_i - \sum_{j=0}^{p} \beta_j X_{ij})^2 = (Y_i - X_i\beta)^2$. We thus want to find the distribution minimizing the expected square-loss, i.e. *squared error risk* (henceforth abbreviated to 'square-risk') relative to the underlying P^*:

$$\text{RISK}^{\text{sq}}(p, \beta) := \mathbf{E}_{(X,Y)\sim P^*}(Y - \mathbf{E}_{p,\beta,\sigma^2}[Y \mid X])^2 = \mathbf{E}_{(X,Y)\sim P^*}(Y - \sum_{j=0}^{p} \beta_j X_j)^2, \tag{3.3}$$

where $\mathbf{E}_{p,\beta,\sigma^2}[Y \mid X]$ abbreviates $\mathbf{E}_{Y\sim P_{p,\beta,\sigma^2}|X}[Y]$. Since this quantity is independent of the variance σ^2, σ^2 is not used as an argument of RISK^{sq}.

3.2.3 KL-associated prediction tasks for the linear model

Suppose that an optimal $\tilde{P} \in \mathcal{M}$ exists in the regression model. We denote by $\tilde{p}$ the smallest p such that $\tilde{P} \in \mathcal{M}_p$, and define $\tilde{\sigma}^2, \tilde{\beta}$ such that $\tilde{P} = P_{\tilde{p},\tilde{\beta},\tilde{\sigma}^2}$. A straightforward computation shows that for all $(p, \beta, \sigma^2) \in \Theta$:

$$\text{RISK}^{\log}((p, \beta, \sigma^2)) = \frac{1}{2\sigma^2}\text{RISK}^{\text{sq}}((p, \beta)) + \frac{1}{2}\log(2\pi\sigma^2), \tag{3.4}$$

so that the (p, β) achieving minimum log-risk for each fixed σ^2 is equal to the (p, β) with the minimum square-risk. In particular, $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ must minimize not just log-risk, but also square-risk. Moreover, the conditional expectation $\mathbf{E}_{P^*}[Y \mid X]$ is known as the *true regression function*. It minimizes the square-risk among all conditional distributions for $Y \mid X$. Together with (3.4) this implies that, if there is some (p, β) such that $\mathbf{E}[Y \mid X] = \sum_{j=0}^{p} \beta_j X_j = X\beta$, i.e. (p, β) represents the true regression function, then $(\tilde{p}, \tilde{\beta})$ also represents the true regression function. In all our examples, this will be the case: the model is misspecified only in that the true noise is heteroskedastic; but the model does invariably contain the true regression function.

Moreover, for each fixed (p, β), the σ^2 minimizing $\text{RISK}^{\log}$ is, as follows by differentiation, given by $\sigma^2 = \text{RISK}^{\text{sq}}(p, \beta)$. In particular, this implies that

$$\tilde{\sigma}^2 = \text{RISK}^{\text{sq}}(\tilde{p}, \tilde{\beta}), \tag{3.5}$$

or in words: the KL-optimal model variance $\tilde{\sigma}^2$ is equal to the true expected (marginal, not conditioned on X) square-risk obtained if one predicts with the optimal $(\tilde{p}, \tilde{\beta})$. This means that the optimal $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ is *reliable* in the sense of Grünwald (1998, 1999): its self-assessment about its square-loss performance is correct, independently of whether $\tilde{\beta}$ is equal to the true regression function or not. In other words, $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ *correctly predicts how well it predicts*.

Summarizing, for misspecified models, $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ is optimal not just in KL/log-risk sense, but also in terms of square-risk and in terms of reliability; in our examples, it also represents the true regression function. We say that, for linear models, square-risk optimality, square-risk reliability and regression-function consistency are *KL-associated prediction tasks*: if (as we hope Bayes will do, but as we will see sometimes does not) we can find the KL-optimal $\tilde{\theta}$, we automatically behave well in these associated tasks as well.

3.3 The generalized posterior

General losses The original generalized posterior is a concept going back at least to Vovk (1990) and has been developed mainly within the so-called (frequentist) *PAC-Bayesian* framework (McAllester, 2003; Seeger, 2002; Catoni, 2007; Audibert, 2004; Zhang, 2006b; see also Bissiri et al. (2013) and the discussion in Section 4.3). It is defined relative to a prior on *predictors* rather than probability distributions. Depending on the decision problem at hand, predictors can be e.g. classifiers, regression functions or probability densities. Formally, we are given an abstract space of predictors represented by a set Θ, which obtains its meaning in terms of a loss function $\ell : \mathcal{Z} \times \Theta \to \mathbf{R}$, writing $\ell_\theta(z)$ as shorthand for $\ell(z, \theta)$. Following e.g. Zhang (2006b), for any prior Π on Θ with density π relative to some underlying measure ρ, we define the *generalized Bayesian posterior with learning rate η relative to loss function ℓ*, denoted as $\Pi \mid Z^n, \eta$, as the distribution on Θ with density

$$\pi(\theta \mid z^n, \eta) := \frac{e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)} \pi(\theta)}{\int e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)} \pi(\theta) \rho(d\theta)} = \frac{e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)} \pi(\theta)}{\mathbf{E}_{\theta \sim \Pi}[e^{-\eta \sum_{i=1}^n \ell_\theta(z_i)}]}. \tag{3.6}$$

Thus, if θ_1 fits the data better than θ_2 by a difference of ϵ according to loss function ℓ, then their posterior ratio is larger than their prior ratio by an amount exponential in ϵ, where the larger η, the larger the influence of the data as compared to the prior.

If $z_i = (x_i, y_i)$ with $y_i \in \mathbf{R}$ and $x_i = (1, x_{i1}, \ldots, x_{ip})$, and the goal is to predict y_i given x_i, then we may take as our prediction model e.g. the set of linear predictors that predict y_i by $\sum \beta_j x_{ij} = x_i \beta$, and as our loss function the squared error loss, $\ell_\beta(x_i, y_i) = (y_i - x_i \beta)^2$. We may then study the behaviour of such a procedure in its own right, irrespective of a Bayesian misspecification interpretation; the experiments we perform in Section 5.1.1 can be interpreted in this manner.

Log-loss and likelihood Now if the set Θ represents a model of (conditional) distributions $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$, we may set, for $z_i = (x_i, y_i)$, $\ell_\theta(z_i) = -\log f_\theta(y_i \mid x_i)$ to be the log-loss as defined above. In this special case, the definition of η-generalized posterior specializes to the definition of 'generalized posterior' as known within the Bayesian literature (Walker and Hjort, 2002; Zhang, 2006a):

$$\pi(\theta \mid z^n, \eta) = \frac{(f(y^n \mid x^n, \theta))^\eta \pi(\theta)}{\int (f(y^n \mid x^n, \theta))^\eta \pi(\theta) \rho(d\theta)} = \frac{(f(y^n \mid x^n, \theta))^\eta \pi(\theta)}{\mathbf{E}_{\theta \sim \Pi}[(f(y^n \mid x^n, \theta))^\eta]}. \quad (3.7)$$

Again, the larger η, the larger the influence of the likelihood. Obviously $\eta = 1$ corresponds to standard Bayesian inference, whereas if $\eta = 0$ the posterior is equal to the prior and nothing is ever learned. Our algorithm for learning η will usually end up with values in between. It has long been known that in model selection and nonparametric settings, there is an issue with consistency proofs for full Bayes, Bayes MAP and MDL if we take the standard $\eta = 1$, and indeed, this is part of the reason why the generalized posterior in the form (3.7) was derived in the first place: for example, Barron and Cover (1991) give general consistency theorems for 2-part MDL (closely related to Bayes MAP) and note that they hold for any $\eta < 1$; but for $\eta = 1$, additional assumptions must be made. Zhang (2006a) gives an explicit example in which the posterior shows anomalous behaviour at $\eta = 1$. A connection to misspecification was first made by Grünwald (2011) (see Section 4.3.1) and Grünwald (2012).

Generalized predictive distribution We also define the predictive distribution based on the η-generalized posterior (3.7) as a generalization of the standard definition as follows: for $m \geq 0, m' \geq m$, we set

$$\begin{aligned} &\bar{f}(y_i, \ldots, y_{i+m} \mid x_i, \ldots, x_{i+m'}, z^{i-1}, \eta) \\ &\quad := \mathbf{E}_{\theta \sim \Pi \mid z^{i-1}, \eta}[f(y_i, \ldots, y_{i+m} \mid x_i, \ldots, x_{i+m'}, \theta)] \\ &\quad = \mathbf{E}_{\theta \sim \Pi \mid z^{i-1}, \eta}[f(y_i, \ldots, y_{i+m} \mid x_i, \ldots, x_{i+m}, \theta)]. \end{aligned} \quad (3.8)$$

where the first equality is a definition and the second follows by our i.i.d. assumption. We always use the bar-notation $\bar{f}$ to indicate marginal and predictive distributions, i.e. distributions on data that are arrived at by integrating out parameters. If $\eta = 1$ then $\bar{f}$ and π become the standard Bayesian predictive density and posterior, and if it is clear from the context that we consider $\eta = 1$, we leave out the η in the notation.

> The generalized posterior is created by exponentiating the likelihood according to individual elements $\theta \in \Theta = \bigcup_p \Theta_p$ in the model and renormalizing, which is not the same as exponentiating marginal likelihoods and renormalizing. In particular, $\pi(p \mid z^n, \eta)$ as given by (3.10) is in general *not* proportional to $(\bar{f}(y^n \mid x^n, p))^\eta \pi(p)$. Similarly, for generalized marginal distributions, as soon as $\eta \neq 1$, we have that in general
>
> $$\bar{f}(y_i, y_{i+1} \mid x_i, x_{i+1}, z^{i-1}, \eta) \neq \bar{f}(y_i \mid x_i, z^{i-1}, \eta) \cdot \bar{f}(y_{i+1} \mid x_{i+1}, z^i, \eta),$$

unlike for the standard Bayesian marginal distribution for which equality holds (in Section 4.2 we encounter a further modification of the generalized posterior whose marginals do satisfy this product rule).

3.3.1 Instantiation to linear model selection and averaging

Now consider again a linear model $\mathcal{M}_p$ as defined in Section 3.2.3. We instantiate the generalized posterior and its marginals for this model. With prior $\pi(\beta, \sigma^2 \mid p)$ taken relative to Lebesgue measure, (3.7) specializes to:

$$\pi(\beta, \sigma \mid z^n, p, \eta) = \frac{(2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)}{\int (2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)\, d\beta\, d\sigma}.$$

Note that in the numerator $1/\sigma^2$ and η are interchangeable in the exponent, but not in the factor in front: their role is subtly different. For Bayesian inference with a sequence of models $\mathcal{M} = \bigcup_{p=0,\ldots,p_{\max}} \mathcal{M}_p$, with $\pi(p)$ a probability mass function on $p \in \{0, \ldots, p_{\max}\}$, we get:

$$\begin{aligned}\pi(\theta \mid z^n, \eta) &= \frac{f(y^n \mid x^n, \theta)^\eta \pi(\theta)}{\int_{\theta\in\Theta} f(y^n \mid x^n, \theta)^\eta \pi(\theta) \rho(d\theta)} \quad \text{with} \quad \theta = (\beta, \sigma^2, p) \\ &= \frac{(2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)\pi(p)}{\sum_{p=0}^{p_{\max}} \int (2\pi\sigma^2)^{-n\eta/2} e^{-\frac{\eta}{2\sigma^2}\sum_{i=1}^n (y_i - x_i\beta)^2} \pi(\beta, \sigma \mid p)\pi(p)\, d\beta\, d\sigma}\end{aligned} \tag{3.9}$$

The total generalized posterior probability of model $\mathcal{M}_p$ then becomes:

$$\pi(p \mid z^n, \eta) = \int \pi(\beta, \sigma, p \mid z^n, \eta)\, d\beta\, d\sigma. \tag{3.10}$$

Analogously to (3.8), for given p, we define the η-generalized Bayesian predictive distribution as:

$$\begin{aligned}\bar{f}(y_i^{i+m} \mid x_i^{i+m'}, z^{i-1}, p, \eta) &:= \mathbf{E}_{\beta,\sigma^2\sim\Pi|z^{i-1},p,\eta}[f(y_i^{i+m} \mid x_i^{i+m'}, \beta, \sigma^2, p)] \\ &= \mathbf{E}_{\beta,\sigma^2\sim\Pi|z^{i-1},p,\eta}[f(y_i^{i+m} \mid x_i^{i+m}, \beta, \sigma^2, p)]\end{aligned} \tag{3.11}$$

(writing a_i^j as shorthand for $a_i, \ldots, a_j$). The previous displays held for general priors. The experiments in this chapter adopt widely used priors (see e.g. Raftery et al., 1997): normal priors on the β's and inverse gamma priors on the variance. These conjugate priors allow explicit analytical formulas for all relevant quantities for arbitrary η, provided below. We only consider the simple case of a fixed $\mathcal{M}_p$ here; the more complicated formulas with an additional prior on p will be given in Appendix 4.A.1 in the next chapter.

Fixed p and σ^2 Let $\mathbf{X}_n = (x_1^\top, \ldots, x_n^\top)^\top$ be the design matrix. For a linear model $\mathcal{M}_p$ with fixed variance σ^2 and initial Gaussian prior on β given by $N(\bar{\beta}_0, \sigma^2\Sigma_0)$, the generalized posterior on β is again Gaussian with mean

$$\bar{\beta}_{n,\eta} := \mathbf{E}_{\beta\sim\Pi|z^n,p,\eta}\, \beta = \Sigma_{n,\eta}(\Sigma_0^{-1}\bar{\beta}_0 + \eta\mathbf{X}_n^\top y^n) \tag{3.12}$$

and covariance matrix $\sigma^2\Sigma_{n,\eta}$, where $\Sigma_{n,\eta} = (\Sigma_0^{-1} + \eta \mathbf{X}_n^\top \mathbf{X}_n)^{-1}$.

Fixed p, varying σ^2 Now consider linear models with a Gaussian prior on β conditional on σ^2 as above, and a conjugate (inverse gamma) prior on σ^2, i.e. $\pi(\sigma^2) = \text{Inv-gamma}(\sigma^2 \mid a_0, b_0)$ for some a_0 and b_0. Here we use the following parameterization of the inverse gamma distribution:

$$\text{Inv-gamma}(\sigma^2 \mid a, b) = \sigma^{-2(a+1)} e^{-b/\sigma^2} b^a / \Gamma(a). \tag{3.13}$$

The posterior $\pi(\sigma^2, z^n, p)$ is then given by $\text{Inv-gamma}(\sigma^2 \mid a_{n,\eta}, b_{n,\eta})$ where

$$a_{n,\eta} = a_0 + \eta n/2 \ ; \ b_{n,\eta} = b_0 + \frac{\eta}{2} \sum_{i=1}^{n} (y_i - x_i \bar{\beta}_{n,\eta})^2. \tag{3.14}$$

The posterior expectation of σ^2 can be calculated as

$$\bar{\sigma}^2_{n,\eta} := \frac{b_{n,\eta}}{a_{n,\eta} - 1}. \tag{3.15}$$

Note that the posterior mean of β given σ^2 does not depend on σ^2.

3.4 The SafeBayesian algorithm

3.4.1 Introducing SafeBayes via the prequential view

We introduce SafeBayes via Dawid's prequential interpretation of Bayes factor model selection. As was first noticed by Dawid (1984) and Rissanen (1984), we can think of Bayes factor model selection as picking the model with index p that, when used for sequential prediction with a logarithmic scoring rule, minimizes the cumulative loss. To see this, note that for any distribution whatsoever, we have that, by definition of conditional probability,

$$-\log f(y^n) = -\log \prod_{i=1}^{n} f(y_i \mid y^{i-1}) = \sum_{i=1}^{n} -\log f(y_i \mid y^{i-1}).$$

In particular, for the standard Bayesian marginal distribution $\bar{f}(\cdot \mid p) = \bar{f}(\cdot \mid p, \eta = 1)$ as defined above, for each fixed p, we have

$$-\log \bar{f}(y^n \mid x^n, p) = \sum_{i=1}^{n} -\log \bar{f}(y_i \mid x^n, y^{i-1}, p) = \sum_{i=1}^{n} -\log \bar{f}(y_i \mid x_i, z^{i-1}, p), \tag{3.16}$$

where the second equality holds by (3.11). If we assume a uniform prior on model index p, then Bayes factor model selection picks the model maximizing $\pi(p \mid z^n)$, which by Bayes' theorem coincides with the model minimizing (3.16), i.e. minimizing cumulative log-loss. Similarly, in 'empirical Bayes' approaches, one picks the value of some nuisance parameter ρ that maximizes

the marginal Bayesian probability $\bar{f}(y^n \mid x^n, \rho)$ of the data. By (3.16), which still holds with p replaced by ρ, this is again equivalent to the ρ minimizing the cumulative log-loss. This is the *prequential* interpretation of Bayes factor model selection and empirical Bayes approaches, showing that Bayesian inference can be interpreted as a sort of *forward* (rather than cross-) validation (Dawid, 1984; Rissanen, 1984; Hjorth, 1982).

We will now see whether we can use this approach with ρ in the role of the η for the η-generalized posterior that we want to learn from the data. We continue to rewrite (3.16) as follows (with ρ instead of p that can either stand for a continuous-valued parameter or for a model index but not yet for η), using the fact that the Bayes predictive distribution given ρ and z^{i-1} can be rewritten as a posterior-weighted average of f_θ:

$$\begin{aligned}\check{\rho} &:= \arg\max_\rho \bar{f}(y^n \mid x^n, \rho) = \arg\min_\rho \sum_{i=1}^n \left(-\log \bar{f}(y_i \mid x_i, z^{i-1}, \rho)\right) \\ &= \arg\min_\rho \sum_{i=1}^n \left(-\log \mathbf{E}_{\theta\sim\Pi|z^{i-1},\rho}[f(y_i \mid x_i, \theta)]\right). \end{aligned} \tag{3.17}$$

This choice for $\check{\rho}$ being entirely consistent with the Bayesian approach, our first idea is to choose $\hat{\eta}$ in the same way: we simply pick the η achieving (3.17), with ρ substituted by η. However as Figure 4.5 will show (the blue line there depicts (3.17) for one of our experiments), this will tend to pick η close to 1 and does not improve predictions under misspecification. Indeed, we introduced η to deal with the case in which the Bayesian model assumptions are violated, so we cannot expect that learning it in a Bayes-like way such as (3.17) will resolve the issue. But it turns out that a *slight* modification of (3.17) does the trick: we simply interchange the order of logarithm and expectation in (3.17) and pick the η minimizing

$$\sum_{i=1}^n \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[-\log f(y_i \mid x_i, \theta)\right]. \tag{3.18}$$

In words, we pick the η minimizing the *P*osterior-*E*xpected *P*osterior-*R*andomized log-loss, i.e. the log-loss we expect to obtain, according to the η-generalized posterior, if we actually sample from this posterior. This modified loss function has also been called *Gibbs error* (Cuong et al., 2013), and while the abbreviation *PEPR*-log-loss would be more correct, we simply call it the η-*R*-log-loss from now on.

A detailed explanation of why this works will be given in Sections 4.1.3 and 4.2; for now we just notice that by Jensen's inequality, for any fixed η, for every sequence of data we must have

$$\mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[-\log f(y_i \mid x_i, \theta)\right] \geq -\log \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[f(y_i \mid x_i, \theta)\right], \tag{3.19}$$

yet, the difference between both sides is small if the posterior is *concentrated* for (x_i, y_i), i.e. for small ϵ and small positive δ, it puts $1-\delta$ of its mass on distributions which assign the same density to y_i given x_i up to a factor $1+\epsilon$ — clearly,

if $\delta = \epsilon = 0$ then both sides are the same. Thus, at values for η at which the generalized posterior is 'cumulatively concentrated', i.e. concentrated at most sample points, the objective function will be similar to the standard Bayesian one. This is the clue to further analysis of the algorithm to follow later.

In practice, it is computationally infeasible to try all values of η and we simply have to try out a number of values. For convenience we give a detailed description of the resulting algorithm below, copied from Grünwald (2012). In this chapter, we will invariably apply it with $z_i = (x_i, y_i)$ as before, and $\ell_\theta(z_i)$ set to the (conditional) log-loss as defined before, although it sometimes also has a second interpretation with ℓ_θ as square-loss.

Algorithm 3.1: The (R-)SafeBayesian algorithm

Input: data $z_1, \ldots, z_n$, model $\mathcal{M} = \{f(\cdot \mid \theta) \mid \theta \in \Theta\}$, prior Π on Θ, step-size κ_{STEP}, max. exponent $\kappa_{\max}$, loss function $\ell_\theta(z)$

Output: Learning rate $\hat{\eta}$

$\mathcal{S}_n := \{1, 2^{-\kappa_{\text{STEP}}}, 2^{-2\kappa_{\text{STEP}}}, 2^{-3\kappa_{\text{STEP}}}, \ldots, 2^{-\kappa_{\max}}\}$;

for all $\eta \in \mathcal{S}_n$ **do**

  $s_\eta := 0$;

  for $i = 1 \ldots n$ **do**

    Determine generalized posterior $\Pi(\cdot \mid z^{i-1}, \eta)$ of Bayes with learning rate η.

    Calculate "posterior-expected posterior-randomized loss" of predicting actual next outcome:

$$r := \ell_{\Pi|z^{i-1},\eta}(z_i) = \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}\left[\ell_\theta(z_i)\right] \tag{3.20}$$

    $s_\eta := s_\eta + r$;

  end

end

Choose $\hat{\eta} := \arg\min_{\eta\in\mathcal{S}_n}\{s_\eta\}$ (if min achieved for several $\eta \in \mathcal{S}_n$, pick largest);

Variation As we will see in Section 4.1.4, the crucial property to make inference about η work is that the expression inside the sum in (3.17) is replaced by

$$\mathbf{E}_{\theta\sim\Pi'}[-\log f_\theta(Y_i \mid X_i)], \tag{3.21}$$

where Π' should be chosen such that the resulting log-loss is as small as possible. In (3.18) we set $\Pi' = \Pi$, but Π' is allowed to be *any* distribution on θ under which the expected log-loss is small. The heuristic analysis of Section 4.1.4 suggests that the smaller the loss that can be formed this way (see also the open problems, Section 4.3.3), the better the resulting method is expected to work.

Now the *η-in-model-log-loss* (or just *η-I-log-loss*), defined as

$$\sum_{i=1}^{n} \left[-\log f(y_i \mid x_i, \mathbf{E}_{\theta \sim \Pi | z^{i-1}, \eta}[\theta]) \right], \tag{3.22}$$

is (by Jensen's inequality) always smaller than (3.18) for the linear models that we consider. This means that, instead of finding the η minimizing (3.18), we may want to find the η minimizing (3.22), which is of the form (3.21) with Π' equal to a point mass on $\bar{\theta}_{i,\eta} := \mathbf{E}_{\theta \sim \Pi | z^{i-1}, \eta} f_\theta$. We call the version of SafeBayes which minimizes the alternative objective function (3.22) *in-model SafeBayes*, abbreviated to *I*-SafeBayes, and from now on use *R*-SafeBayes for the original version based on the *R*-log-loss. We did not realize the potential benefits of using in-model SafeBayes at the time of writing Grünwald (2012), and while the theoretical results of Grünwald can be adjusted to deal with such modifications, we cannot get any better theoretical convergence bounds as yet, but this may be an artefact of our proof techniques. A secondary goal of the experiments in this chapter is thus to see whether one can really improve SafeBayes by using the 'in-model' version.

3.4.2 Instantiating SafeBayes to the linear model

Our experiments concern four instantiations of SafeBayes: *R*-SafeBayes and *I*-SafeBayes for models with fixed variance, denoted *R-square-SafeBayes and I-square-SafeBayes* for reasons that will become clear below, are the topic of experiments in Section 5.1.1. The main text instead investigates, in Section 3.5, *R*-SafeBayes and *I*-SafeBayes for models with varying variance, denoted *R-log-SafeBayes and I-log-SafeBayes*. Below we give explicit formulas for each when conditioned on a fixed model $\mathcal{M}_p$; the case with a posterior on p itself can easily be derived from these.

Fixed σ^2: *R*-square- and *I*-square-SafeBayes When conditioned on a fixed p and σ^2 (a situation with which we experiment in Section 5.1.1.2), SafeBayes tries to minimize the *R*-log-loss, which, as an easy calculation shows, is just the sum, from $i = 0$ to $n-1$, of

$$\begin{aligned} &\mathbf{E}_{\beta \sim \Pi | z^i, p, \eta} \left[-\log f(y_{i+1} \mid x_{i+1}, \beta, \sigma^2) \right] \\ &\qquad = \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} (y_{i+1} - x_{i+1} \bar{\beta}_{i,\eta})^2 + \frac{1}{2} x_{i+1} \Sigma_{i,\eta} x_{i+1}^\top, \end{aligned} \tag{3.23}$$

where $\bar{\beta}_{i,\eta}$ and $\Sigma_{i,\eta}$ are given as in and below (3.12). Note that $\bar{\beta}_{i,\eta}$ depends on η but not on σ, and note also that, since $\mathbf{X}_n^\top \mathbf{X}_n$ (as in (3.12)) tends to increase linearly in n and p, the final term is of order $p/(n\eta)$.

In the corresponding in-model version of SafeBayes, we use the in-model-loss as given by $-\log f(y_{i+1} \mid x_{i+1}, \bar{\beta}_{i,\eta}, \sigma^2)$, which is equal to (3.23) without the final term. Since the first term of (3.23) does not depend on the data, this version of SafeBayes thus amounts to picking the $\hat{\eta}$ minimizing just the sum

of square-loss prediction errors, *which does not depend on the chosen* σ^2. It thus becomes a standard version of 'prequential model selection' as based on the square-loss, which in turn is similar to (though having different asymptotics than) leave-one-out cross validation based on the square-loss.

Indeed, the fixed σ^2 versions of SafeBayes can be interpreted in two ways: first, as we did until now, in terms of SafeBayes with ℓ_θ in (3.20) set to the log-loss, i.e. as a tool for dealing with misspecification; and second, with ℓ_θ in (3.20) set proportionally to the square-loss, as a generic tool to learn good square-loss predictors (not distributions) in a pseudo-Bayesian way. More precisely, I-SafeBayes with the log-loss for fixed σ^2 is equivalent to the version of I-SafeBayes we would get if we set $\ell_{\beta,\sigma^2}(x,y) := C(y - x\beta)^2$, for any constant $C > 0$. Similarly, R-SafeBayes with the log-loss for fixed σ^2 is equivalent to the version of R-SafeBayes we would get if we set $\ell_{\beta,\sigma^2}(x,y) := C(y - x\beta)^2$, although now equivalence only holds if we set $C = 1/2\sigma^2$. For this reason we will now refer to them as *I-square-SafeBayes and R-square-SafeBayes*, respectively.

Varying σ^2: R-log- and I-log-SafeBayes Next consider the situation with fixed p and varying σ^2, with posterior on σ^2 an inverse gamma distribution with parameters $a_{n,\eta}$ and $b_{n,\eta}$ as given by (3.14). Then the R-log-loss is given by

$$\begin{aligned}
&\mathbf{E}_{\sigma^2,\beta\sim\Pi|z^i,p,\eta}\left[-\log f(y_{i+1} \mid x_{i+1},\beta,\sigma^2)\right] \\
&\quad= \frac{1}{2}\log 2\pi b_{i,\eta} - \frac{1}{2}\psi(a_{i,\eta}) + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2}{b_{i,\eta}/a_{i,\eta}} + \frac{1}{2}x_{i+1}\Sigma_{i,\eta}x_{i+1}^\top \\
&\quad= \frac{1}{2}\log 2\pi\bar{\sigma}^2_{i,\eta} + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2}{\bar{\sigma}^2_{i,\eta}} + \frac{1}{2}x_{i+1}\Sigma_{i,\eta}x_{i+1}^\top + r(i,\eta), \quad (3.24)
\end{aligned}$$

where ψ is the digamma function, $\bar{\sigma}^2_{i,\eta}$ is the η-posterior expectation of σ^2 as given by (3.15) and $r(i,\eta)$ is a remainder function which is $O(1/i)$ whenever $\sum_{i=1}^n (y_i - x_i\beta_{n,\eta})^2$ increases linearly in i. This final approximation follows by (3.15) and because we have $\psi(x) \in [\log(x-1), \log x]$. R-SafeBayes for varying σ^2 minimizes (3.24), and, because there is now only a log-loss and not a direct square-loss interpretation, we will call it *R-log-SafeBayes* from now on.

To calculate the corresponding in-model version of SafeBayes, *I-log-SafeBayes*, note that it minimizes the sum of

$$-\log f(y_{i+1} \mid x_{i+1},\bar{\beta}_{i,\eta},\bar{\sigma}^2_{i,\eta}) = \frac{1}{2}\log 2\pi\bar{\sigma}^2_{i,\eta} + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\beta_{i,\eta})^2}{\bar{\sigma}^2_{i,\eta}}. \quad (3.25)$$

Comparing the four versions of SafeBayes, we see that both R-SafeBayeses have an additional term which decreases in η, increases in model dimensionality p (via the size of the matrix $\Sigma_{i,\eta}$), but becomes negligible for $n \gg p$.

3.4.3 SafeBayes learns to predict as well as the optimal distribution

We first define the *Cesàro-averaged* posterior given data Z^n by setting, for any subset $\Theta' \subset \Theta$,

$$\Pi_{\text{Ces}}(\Theta' \mid Z^n, \eta) := \frac{1}{n} \sum_{i=1}^{n} \Pi(\Theta' \mid Z^i, \eta) \tag{3.26}$$

to be the posterior probability of Θ' averaged over the n posterior distributions obtained so far. Predicting based on Cesàro-averaged posteriors was introduced independently by several authors (Barron, 1987; Helmbold and Warmuth, 1992; Yang, 2000; Catoni, 1997) and has received a lot of attention in the machine learning literature in recent years, also under the name "on-line to batch conversion of Bayes" or *progressive mixture rule* (Audibert, 2007) or *mirror averaging* (Juditsky et al., 2008; Dalalyan and Tsybakov, 2012), but is of course unnatural from a Bayesian perspective.

The main result of Grünwald (2012) essentially states the following: suppose that, under P^*, the density ratios are uniformly bounded, i.e. there is a finite v such that for all $\theta, \theta' \in \Theta$, $P^*(f_\theta(Y \mid X)/f_{\theta'}(Y \mid X) \leq v) = 1$. Suppose further that the prior Π assigns 'sufficient mass' in KL-neighbourhoods of $P_{\tilde{\theta}}$. Then Π_{Ces} applied with the $\hat{\eta}$ learned by the SafeBayesian algorithm concentrates on the optimal $P_{\tilde{\theta}}$. That is, let Θ_δ be the subset of all $\theta \in \Theta$ with $D(P^* \| P_\theta) \geq D(P^* \| P_{\tilde{\theta}}) + \delta$. Then for all $\delta > 0$, with P^*-probability 1, as $n \to \infty$, we have that $\Pi_{\text{Ces}}(\Theta_\delta \mid Z^n, \hat{\eta})$ goes to 0. Grünwald goes on to show that in several settings, one can design priors such that the rate at which the posterior concentrates is minimax optimal, i.e. no algorithm can do better in general. On the negative side, the requirement of bounded density ratio is strong, and the replacement of the standard posterior by the Cesàro one is awkward. On the positive side, the theorem has no further conditions and can be applied to parametric and nonparametric cases alike.

In recent, as yet unpublished work, Grünwald (2014) extends the result to deal with unbounded density ratios as in the regression setting considered here, and to the 'standard' η-generalized rather than the Cesàro-averaged η-generalized posterior. In both cases, convergence can still be proved but the bounds given on the concentration rate worsen by a $\log n$ factor. We suspect that in many situations, this is an artefact of the proof technique, and to see whether there is any practical difference, below we include experimental results both for the Cesàro-averaged η-generalized posterior $\Pi_{\text{Ces}}(\cdot \mid Z^n, \hat{\eta})$ and for the standard η-generalized posterior $\Pi(\cdot \mid Z^n, \hat{\eta})$.

3.5 Main experiment: Varying σ^2

In this section we provide our main experimental results, based on linear models $\mathcal{M}_p$ as defined in Section 3.2.2 with a prior on both the mean and the variance. Figures 3.3–3.6 depict, and Section 3.5.3 discusses the results of model selection and averaging experiments, which choose or average between the

models $0, \ldots, p_{\max}$, where we consider first an incorrectly and then a correctly specified model, both with $p_{\max} = 50$ and later with $p_{\max} = 100$. Section 3.5.4 contains and interprets additional experiments on Bayesian ridge regression, with a fixed p; a multitude of additional experiments is provided in Chapter 5. Section 3.5.5 at the end of this chapter summarizes the relevant findings of these additional experiments.

3.5.1 Preparing the main experiments: Model, priors, method, 'truth'

In this subsection we prepare the experiments: Section 3.5.1.1 describes our priors π; Section 3.5.1.2 concerns the sampling ('true') distributions P^* with which we experiment; and finally, Section 3.5.2 describes the data statistics that we will report.

3.5.1.1 The priors

Prior on models In our model selection/averaging experiments, we use a fat-tailed prior on the models given by

$$\pi(p) \propto \frac{1}{(p+2)(\log(p+2))^2}.$$

This prior was chosen because it remains well-defined for an infinite collection of models, even though we only use finitely many in our experiments.

Variation As a sanity check we did repeat some of our experiments with a uniform prior on $0, \ldots, p_{\max}$ instead; the results were indistinguishable.

Prior on parameters given models Each model $\mathcal{M}_p$ has parameters β, σ^2, on which we put the standard conjugate priors as described in Section 3.3.1. We set the mean of the prior on β to $\bar{\beta}_0 = \mathbf{0}$, and its covariance matrix to $\sigma^2\Sigma_0$. Our main experiments below are based on an *informative* instantiation of Σ_0, using the identity matrix $\Sigma_0 = \mathbf{I}_{p+1}$; this prior equals the posterior we would get by starting with an improper Jeffreys' prior on β and then observing, for each coefficient β_j, one extra point $z = (x, 0)$ with $x_j = 1$ and $x_i = 0$ for $i \neq j$.

Variations We also ran experiments with a 'slightly informative' Σ_0, where we set $\Sigma_0 = 1000 \cdot \mathbf{I}_{p+1}$, comparable to observing points $z = (x, 0)$ with $x_j = 1/\sqrt{1000}$. Finally, following the standard reference Raftery et al. (1997), we also used a prior with a level of informativeness depending on the submodel, described in more detail in Section 5.1.

As to the prior on σ^2: Jeffreys' prior is obtained for the choice $a_0 = b_0 = 0$ in (3.13). We do not use this improper prior, because of the well-known issues with Bayes factors under improper priors (O'Hagan, 1995). Moreover, to calculate the posterior's reliability (defined in Section 3.5.2 and shown in Figure 3.3) and also for the I-log-loss, we need to calculate the posterior expectation of the variance σ^2 quantity as given by (3.15), which is only well-defined and finite for $a_n > 1$. We want to make $\pi(\sigma^2)$ as uninformative as possible while ensuring

that (for any positive learning rate) this variance exists for the posterior based on at least one sample. This is accomplished by choosing $a_0 = 1$: for standard Bayes, the posterior after one observation has $a_1 = a_0 + 1/2$; for generalized Bayes, $a_1 = a_0 + \eta/2$. To set b_0, we use that b_0/a_0 represents the sample variance of a virtual initial data sequence (Gelman et al., 2013, Section 14.8). We choose $b_0 = 1/40$ so that $b_0/a_0 = 1/40$, the true variance of the noise in our data, as we describe next.

3.5.1.2 The "truth" (sampling distribution)

Our experiments fall into two categories: correct-model and wrong-model experiments.

Correct-model experiments Here $X_1, X_2, \ldots$ are sampled i.i.d., with, for each individual $X_i = (X_{i1}, \ldots, X_{ip_{\max}})$, $X_{i1} \ldots, X_{ip_{\max}}$ i.i.d. $\sim N(0,1)$. Given each X_i, Y_i is generated as

$$Y_i = .1 \cdot (X_{i1} + \ldots + X_{i4}) + \epsilon_i, \tag{3.27}$$

where the ϵ_i are i.i.d. $\sim N(0, \sigma^{*2})$ with variance $\sigma^{*2} = 1/40$.

Wrong-model experiments Now at each time point i, a fair coin is tossed independently of everything else. If the coin lands heads, then the point is 'easy', and $(X_i, Y_i) := (\mathbf{0}, 0)$. If the coin lands tails, then X_i is generated as for the correct model, and Y_i is generated as (3.27), but now the noise random variables have variance $\sigma_0^2 = 2\sigma^{*2} = 1/20$. Thus, $Z_i = (X_i, Y_i)$ is generated as in the true model case but with a larger variance; this larger variance has been chosen so that the marginal variance of each Y_i is the same value σ^{*2} in both experiments.

From the results in Section 3.2.3 we immediately see that, for both experiments, the optimal model is $\mathcal{M}_{\tilde{p}}$ for $\tilde{p} = 4$, and the optimal distribution in $\mathcal{M}$ and $\mathcal{M}_{\tilde{p}}$ is parameterized by $\tilde{\theta} = (\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ with $\tilde{p} = 4$, $\tilde{\beta} = (\tilde{\beta}_0, \ldots \tilde{\beta}_4) = (0, .1, .1, .1, .1), \tilde{\sigma}^2 = 1/40$ (in the correct model experiment, $\tilde{\sigma}^2 = \sigma^{*2}$; in the wrong model experiment, since $\tilde{\sigma}^2$ must be reliable, it must be equal to the square-risk obtained with $(\tilde{p}, \tilde{\beta})$, which is $(1/2) \cdot (1/20) = 1/40$). $f(x) := x\tilde{\beta}$ is then equal to the *true* regression function $\mathbf{E}_{P^*}[Y \mid X]$.

Variations We have already seen a variation of these two experiments depicted in Figures 3.1 and 3.2. In the correct-model version of that experiment, P^* is defined as follows: set $X_j = P_j(S)$, where P_j is the Legendre polynomial of degree j and S is uniformly distributed on $[-1, 1]$, and set $Y = 0 + \epsilon$, where $\epsilon \sim N(0, \sigma^{*2})$, with $\sigma^{*2} = 1/40$; $(X_1, Y_1), \ldots$ are then sampled as i.i.d. copies of (X, Y). Note that the true regression function is 0 here. In Section 5.3 we briefly consider this and several other variations of these ground truths.

3.5.2 The statistics we report

Figure 3.3 reports the results of the wrong-model, $p = 50$ experiment; Figure 3.4 shows correct-model, $p = 50$; Figure 3.5 is about wrong-model, $p =$

100; and Figure 3.6 depicts the correct-model, $p = 100$ setting. For all four experiments we measure three aspects of the performance of Bayes and SafeBayes, each summarized in a separate graph. First, we show the behaviour of several prediction methods based on SafeBayes relative to square-risk; second, we measure whether the methods provide a good assessment of their own predictive capabilities in terms of square-loss, i.e. whether they are reliable and not 'overconfident'. Third, we check a form of model identification consistency. Below we explain these three performance measures in detail. We postpone all experiments with log-loss rather than square-loss to Section 4.1.4. We also provide a fourth graph in each case indicating what $\hat{\eta}$'s are typically selected by the two versions of SafeBayes.

Square-risk For a given distribution W on (p, β, σ^2), the *regression function based on* W, a function mapping covariate X to $\mathbf{R}$, abbreviated to $\mathbf{E}_W[Y \mid X]$, is defined as

$$\mathbf{E}_W[Y \mid X] := \mathbf{E}_{(p,\beta,\sigma)\sim W} \mathbf{E}_{Y\sim P_{p,\beta,\sigma}|X}[Y] = \mathbf{E}_{(p,\beta,\sigma)\sim W}\left[\sum_{j=0}^{p} \beta_j X_j\right]. \tag{3.28}$$

If we take W to be the η-generalized posterior, then (3.28) is also simply called the η-posterior regression function. The *square-risk* relative to P^* based on predicting by W is then defined as an extension of (3.3) as

$$\text{RISK}^{\text{sq}}(W) := \mathbf{E}_{(X,Y)\sim P^*}(Y - \mathbf{E}_W[Y \mid X])^2. \tag{3.29}$$

In the experiments below we measure the square-risk relative to P^* at sample size $i-1$ achieved by, respectively, (1), the η-generalized posterior, (2), the η-generalized posterior conditioned on the MAP (maximum a posteriori) model, and, (3), the η-generalized Cesàro-averaged posteriors, i.e.

$$\begin{aligned} &\mathbf{E}_{Z^{i-1}\sim P^*}[\text{RISK}^{\text{sq}}(W)], \text{ with} \\ &W = \Pi \mid Z^{i-1}, \eta \; ; \; W = \Pi \mid Z^{i-1}, \eta, \breve{p}_{\text{map}(Z^{i-1},\eta)} \; ; \; W = \Pi_{\text{CES}} \mid Z^{i-1}, \eta, \end{aligned} \tag{3.30}$$

respectively, where the MAP model $\breve{p}_{\text{map}(Z^{i-1},\eta)}$ is defined as the p achieving $\max_{p\in 0,\ldots,p_{\max}} \pi(p \mid Z^{i-1}, \eta)$, with $\pi(p \mid Z^{i-1}, \eta)$ defined as in (3.10), and Π_{CES} is the Cesàro-averaged posterior as defined as in (3.26). We do this for three values of η: (a) $\eta = 1$, corresponding to the standard Bayesian posterior, (b) $\eta := \hat{\eta}(Z^{i-1})$ set by the R-log SafeBayesian algorithm run on the past data Z^{i-1}, and (c) η set by the I-log SafeBayesian algorithm. In the figures of Section 3.5.3, 1(a) is abbreviated to *Bayes*, 1(b) is *R-log-SafeBayes*, 1(c) is *I-log-SafeBayes*, 2(a) is *Bayes MAP*, 2(b) is *R-log-SafeBayes MAP*, 2(c) is *I-log-SafeBayes MAP*, and results with Cesàro-averaging are discussed but not explicitly shown. In Section 3.5.4, additionally 3(a) is *Bayes Cesàro*, 3(b) is *R-log-SafeBayes Cesàro*, and 3(c) is *I-log-SafeBayes Cesàro*.

Concerning the three square-risks that we record: The first choice is the most natural, corresponding to the prediction (regression function) according to the 'standard' η-generalized posterior; the second corresponds to the

situation where one first selects a single submodel $\check{p}_{\text{map}}$ and then bases all predictions on that model; it has been included because such methods are often adopted in practice. The third choice, the *Cesàro-averaged generalized posterior* is included because, when $\eta = \hat{\eta}$ is set by SafeBayes, this is the choice that Grünwald (2012) provides theoretical convergence results for (as we discussed, Grünwald (2014) provides results for the non-averaged η-generalized posterior as well, but these are worse by a log-factor). But we are also interested in the results for the Cesàro-average for $\eta = 1$, because this has been proposed earlier — albeit somewhat implicitly and with different models — to stabilize Bayesian predictions in adversarial circumstances (Helmbold and Warmuth, 1992), so we include these as well.

In Figure 3.3 and subsequent figures below, we depict these quantities by sequentially sampling data $Z_1, Z_2, \ldots, Z_{\text{max}}$ i.i.d. from a P^* as defined above in Section 3.5.1.2, where max is some large number. At each i, after the first $i-1$ points Z^{i-1} have been sampled, we compute the three square-risks given above. We repeat the whole procedure a number of times (called 'runs'); the graphs show the average risks over these runs.

MAP-model identification / Occam's razor When the goal of inference is model identification, 'consistency' of a method is often defined as its ability to identify the smallest model $\mathcal{M}_{\tilde{p}}$ containing the 'pseudo-truth' $(\tilde{\beta}, \tilde{\sigma}^2)$. To see whether standard Bayes and/or SafeBayes are consistent in this sense, we check whether the MAP model $\check{p}_{\text{map}(Z^{i-1},\eta)}$ is equal to $\tilde{p}$.

Reliability vs. overconfidence Does Bayes learn how good it is in terms of squared error? To answer this question, we define, for a predictive distribution W as in (3.29) above, $U_i^{[W]}$ (a function of X_i, Y_i and (through W) of Z^{i-1}), as

$$U_i^{[W]} = (Y_i - \mathbf{E}_W[Y_i \mid X_i])^2.$$

This is the error we make if we predict Y_i using the regression function based on prediction method W. In the graphs in the next sections we plot the *self-confidence ratio* $\mathbf{E}_{X_i,Y_i\sim P^*}[U_i^{[W]}] / \mathbf{E}_{X_i\sim P^*}\mathbf{E}_{Y_i\sim W|X_i}[U_i^{[W]}]$ as a function of i for the three prediction methods / choices of W defined above. We may think of this as the ratio between the actual expected prediction error (measured in square-loss) one gets by using a predictor who based predictions on W and the marginal (averaged over X) subjectively expected prediction error by this predictor. We previously, in Section 3.2.3, showed that the KL-optimal $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ is *reliable*: this means that, if we would take W the point mass on $(\tilde{p}, \tilde{\beta}, \tilde{\sigma}^2)$ and thus, irrespective of past data Z^{i-1}, would predict by $\mathbf{E}_{(\tilde{p},\tilde{\beta},\tilde{\sigma}^2)}[Y_i \mid X_i] = \sum_{j=0}^{\tilde{p}} \tilde{\beta}_j X_{ij}$, then the ratio would be 1. For the W learned from data considered above, a value larger than 1 indicates that W does not implement a 'reliable' method in the sense of Section 3.2.3, but rather overconfident: it predicts its predictions to be better than they actually are, in terms of square-risk.

3.5.3 Main model selection/averaging experiment

We run the SafeBayesian algorithm of Section 3.4 with $z_i = (x_i, y_i)$ and $\ell_\theta(z_i) = -\log f_\theta(y_i \mid x_i)$ is the (conditional) log-loss as described in that section. As to the parameters of the algorithm (page 52), in all experiments we set the step-size $\kappa_{\text{STEP}} = 1/3$ and $\kappa_{\max} := 8$, i.e. we tried the following values of η: $1, 2^{-1/3}, 2^{-2/3}, \ldots, 2^{-8}$. The result of the wrong-model and correct-model experiment as described above with $p_{\max} = 50$ and $p_{\max} = 100$, respectively, are given in Figures 3.3–3.6.

Conclusion 1: Bayes performs well in model-correct, and dismally in model-incorrect experiment The four figures show that standard Bayes behaves excellently in terms of all quality measures (square-risk, MAP model identification and reliability) when the model is correct, and dismally if the model is incorrect.

Conclusion 2: If (and only if) model incorrect, then the higher $p_{\max}$, the worse Bayes gets We see from Figures 3.4 and 3.6 that standard Bayes behaves excellently in terms of all quality measures (square-risk, MAP model identification and reliability) when the model is correct, both if $p_{\max} = 50$ and if $p_{\max} = 100$, the behaviour at $p_{\max} = 100$ being essentially indistinguishable from the case with $p_{\max} = 50$. These and other (unreported) experiments strongly suggests that, when the data are sampled from a low-dimensional model, then, when the model is correct, standard Bayes is unaffected (does not get confused) by adding additional high-dimensional models to the model space. Indeed, the same is suggested by various existing Bayesian consistency theorems, such as those by Doob (1949); Ghosal et al. (2000); Zhang (2006a).

At the same time, from Figures 3.3 and 3.5 we infer that standard Bayes behaves very badly in all three quality measures in our (admittedly very 'evilly chosen') model-wrong experiment. At very large sample sizes, Bayes eventually recovers, but the main point here to notice is that the n at which a given level of recovery (measured in, say, square-loss) takes place is much higher for the case $p_{\max} = 100$ (Figure 3.5) than for the case $p_{\max} = 50$ (Figure 3.3). This strongly suggests that, when the model is incorrect but the best approximation lies in a low-dimensional submodel, then standard Bayes gets confused by adding additional high-dimensional models to the model space — recovery takes place at a sample size that increases with $p_{\max}$. Indeed, the graphs strongly suggest that in the case that $p_{\max} = \infty$ (with which we cannot experiment), Bayes will be inconsistent in the sense that the risk of the posterior predictive will never ever reach the risk attainable with the best submodel. Grünwald and Langford (2007) showed that this can indeed happen with a simple, but much more unnatural classification model; the present result indicates (but does not prove) that it can happen with our standard model as well.

Conclusion 3: R-log-SafeBayes and I-log-SafeBayes generally perform well Comparing the four graphs for SafeBayes and I-log-SafeBayes, we see that

they behave quite well for *both* the model-correct and the model-wrong experiments, being slightly worse than, though still competitive to, standard Bayes when the model is correct and incomparably better when the model is wrong. Indeed, in the wrong-model experiments, about half of the data points are identical and therefore do not provide very much information, so one would expect that if a 'good' method achieves a given level of square-risk at sample size n in the correct-model experiment, it achieves the same level at about $2n$ in the incorrect-model experiment, and this is indeed what happens. Also, we see from comparing Figures 3.5 and 3.6 on the one hand to Figures 3.3 and 3.4 on the other that adding additional high-dimensional models to the model space hardly affects the results — like standard Bayes when the model is correct, SafeBayes does not get confused by the additional, larger model space.

Secondary conclusions We see that both types of SafeBayes converge quickly to the right (pseudo-true) model order, which is pleasing since they were not specifically designed to achieve this. Whether this is an artefact of our setting or holds more generally would, of course, require further experimentation. We note that at small sample sizes, when both types of SafeBayes still tend to select an overly simple model, I-log-SafeBayes has significantly more variability in the model chosen-on-average; it is not clear though whether this is 'good' or 'bad'. We also note that the η's chosen by both versions are very similar for all but the smallest sample sizes, and are consistently smaller than 1. When instead of the full η-generalized posteriors, the η-generalized posterior conditioned on the MAP $\check{p}_{\text{map}}$ is used, the behaviour of all method consistently deteriorates a little, but never by much.

For lack of space in the graphs, we did not show the Cesàro-versions of Bayes, R-log-SafeBayes and I-log-SafeBayes (methods 3(a), 3(b), 3(c) in Section 3.5.2). Briefly, the curves look as follows: Cesàro-Bayes performs significantly better than standard Bayes in all three quality measures in the wrong-model experiments, but is still far from competitive with the two (full-posterior) SafeBayes versions. When Cesàroified, the SafeBayes methods become a bit smoother but not necessarily better. Very similar behaviour of Cesàro (making bad methods significantly better but still not competitive, and good methods smoother, sometimes a bit worse and sometimes a bit better) has been explicitly depicted in the ridge regression with varying σ^2 in Section 3.5.4 below.

3.5.4 Second experiment: Ridge regression, varying σ^2

We repeat the model-wrong and model-correct experiments of Figures 3.3 and 3.4, with just one major difference: all posteriors are conditioned on $p := p_{\max} = 50$. Thus, we effectively consider just a fixed, high-dimensional model, whereas the best approximation $\tilde{\theta} = (50, \tilde{\beta}, \tilde{\sigma}^2)$ viewed as an element of $\mathcal{M}_p$ is 'sparse' in that it has only $\beta_1, \ldots, \beta_4$ not equal to 0. We note that the MAP model index graphs of Figures 3.3 and 3.4 are meaningless in this context (they

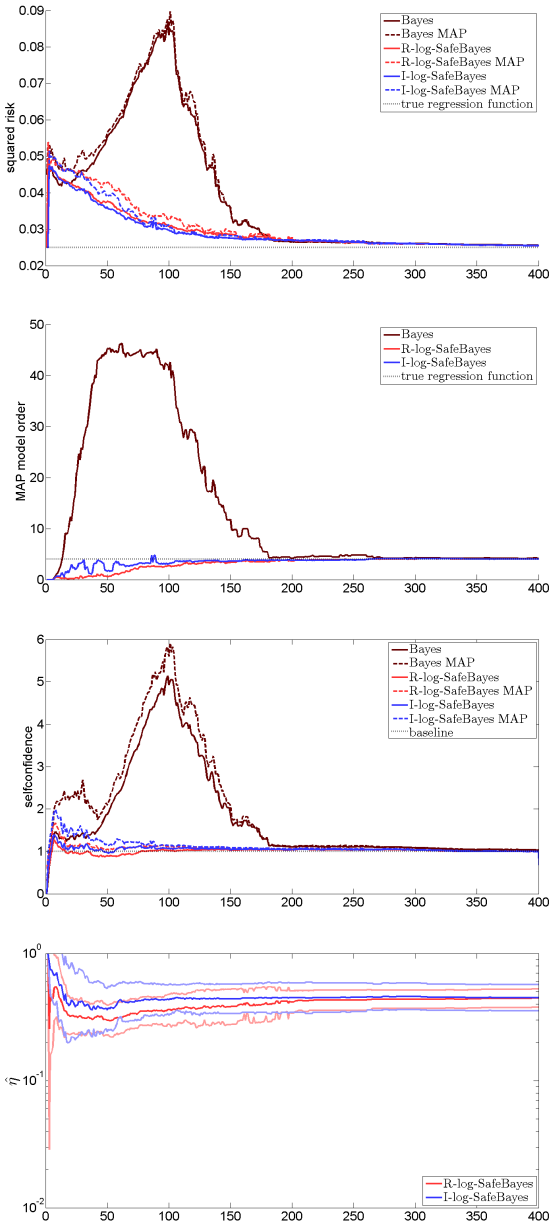

Figure 3.3: Four graphs showing respectively the square-risk, MAP model order, overconfidence (lack of reliability), and selected $\hat{\eta}$ at each sample size, each averaged over 30 runs, for the wrong-model experiment with $p_{\max} = 50$, for the methods indicated in Section 3.5.2. For the selected-$\hat{\eta}$ graph, the pale lines are one standard deviation apart from the average; all lines in this graph were computed over $\hat{\eta}$ indices (so that the lines depict the geometric mean over the values of $\hat{\eta}$ themselves).

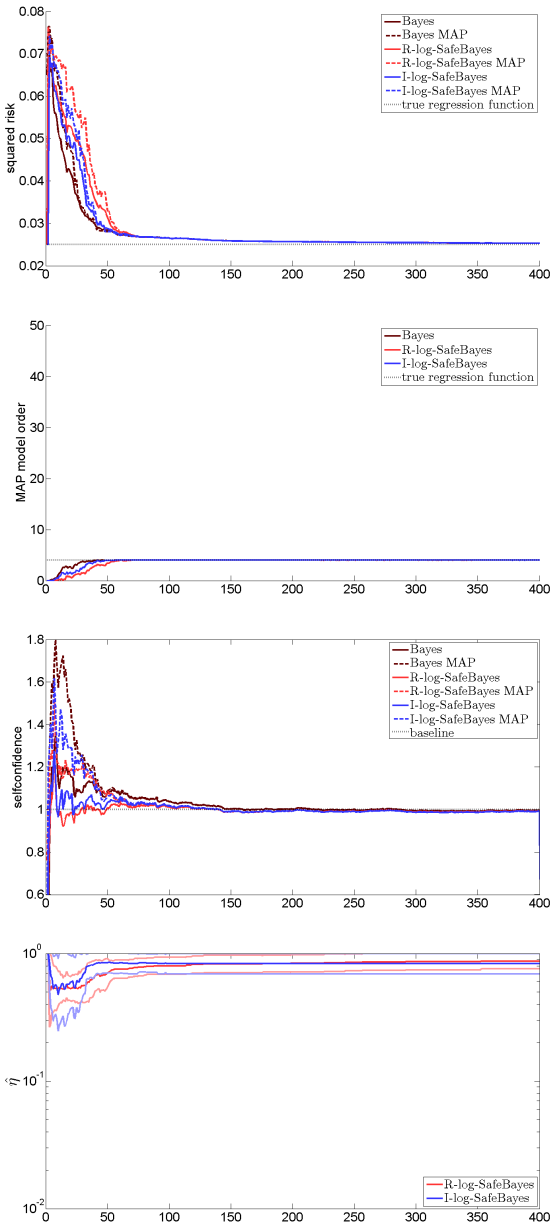

Figure 3.4: Same graphs as in Figure 3.3 for the correct-model experiment with $p_{\max} = 50$

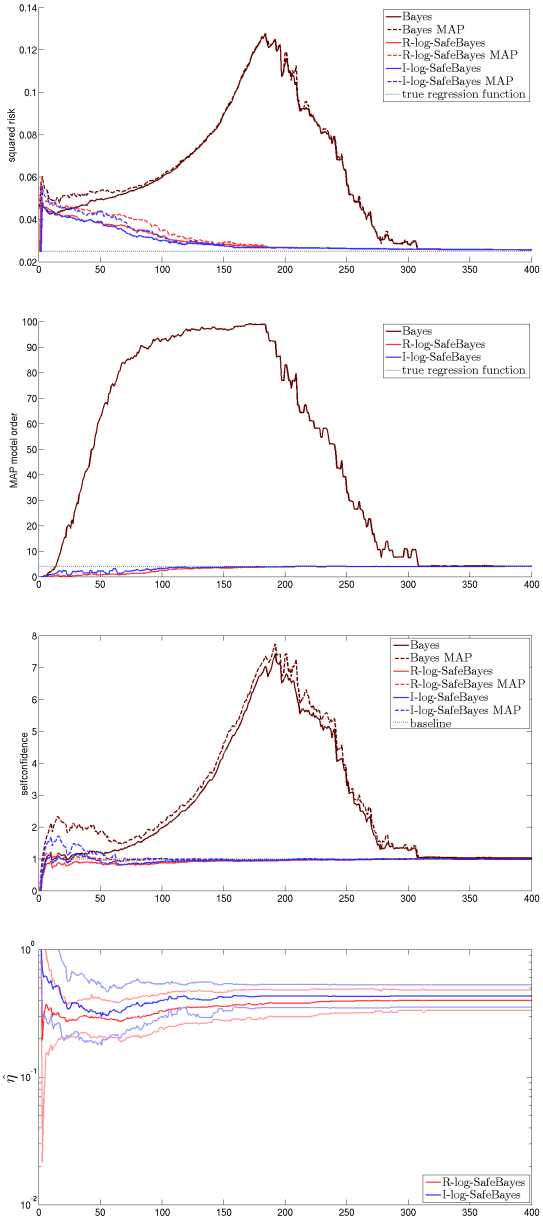

Figure 3.5: Same four graphs as in Figure 3.3, for the wrong-model experiment with $p_{\max} = 100$

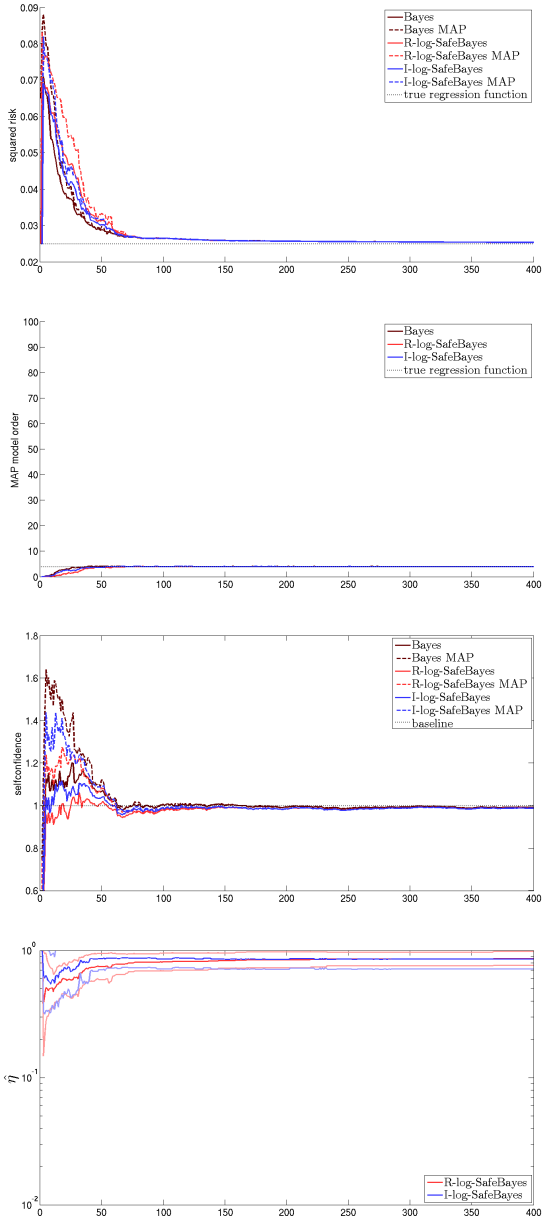

Figure 3.6: Same graphs as in Figure 3.3 for the correct-model experiment with $p_{\max} = 100$

would be equal to the constant 50) so they are left out of the new Figures 3.7 and 3.8.

Instantiating SafeBayes Since we noticed in preliminary experiments that some versions of SafeBayes now have a tendency to select much smaller values of η than in the previous experiments, we now set $\kappa_{\max} = 16$ (large enough so that in no experiment the optimal $\eta < 2^{-\kappa_{\max}}$); for computational reasons we also increased the step size and set $\kappa_{\text{STEP}} = 1$.

Connection to Bayesian (b)ridge regression From (3.12) we see that the posterior mean parameter $\bar{\beta}_{i,\eta}$ is equal to the posterior MAP parameter and depends on η but not on σ^2, since σ^2 enters the prior in the same way as the likelihood. Therefore, the square-loss obtained when using the generalized posterior for prediction is always given by $(y_i - x_i\bar{\beta}_{i,\eta})^2$ irrespective of whether we use the posterior mean, or MAP, or the value of σ^2. Interestingly, if we fix some λ and perform standard (nongeneralized) Bayes with a modified prior, proportional to the original prior raised to the power $\lambda := \eta^{-1}$, then the prior becomes normal $N(\bar{\beta}_0, \sigma^2\Sigma_0')$ where $\Sigma_0' = \eta\Sigma_0$ and the standard posterior given z^i is then (by (3.12)) Gaussian with mean

$$\left(\left(\Sigma_0'\right)^{-1} + \mathbf{X}_n^\top \mathbf{X}\right)^{-1} \cdot \left(\left(\Sigma_0'\right)^{-1} \bar{\beta}_0 + \mathbf{X}_n^\top y^n\right) = \bar{\beta}_{i,\eta}. \tag{3.31}$$

Thus we see that in this special case, the (square-risk of the) η-generalized Bayes posterior mean coincides with the (square-risk of the) standard Bayes posterior mean with prior $N(\bar{\beta}_0, \sigma^2\eta\Sigma_0)$. But this means that the square-loss obtained by η-generalized Bayes on a data sequence is precisely equal to the square-loss obtained by *Bayesian ridge regression* with penalty parameter $\lambda = \eta^{-1}$, as defined, by, e.g., Park and Casella (2008) (to be precise, they call this method Bayesian 'bridge' regression with $q = 2$; the choice of $q = 1$ in their formula gives their celebrated 'Bayesian Lasso'). It is thus of interest to see what happens if η (equivalently, λ) is determined by *empirical Bayes*, which is one of the methods Park and Casella (2008) suggest. In addition to the graphs discussed earlier in Section 3.5.2, we thus also show the results for η set in this alternative way. Whereas this empirical-Bayesian ridge regression is usually a very competitive method (indeed in our model-correct experiment, Figure 3.8, it performs best in all respects), we will see in Figure 3.7 (the green line) that, just like other versions of Bayes, it breaks down under our type of misspecification.

We hasten to add that the correspondence between the η-generalized posterior means and the standard posterior means with prior raised to power $1/\eta$ only holds for the $\bar{\beta}_{i,\eta}$ parameters. It does not hold for the $\bar{\sigma}^2_{i,\eta}$ parameters, and thus, for fixed η, the self-confidence ratio of both methods may be quite different.

Conclusions for model-wrong experiment For most curves, the overall picture of Figure 3.7 is comparable to the corresponding model averaging experi-

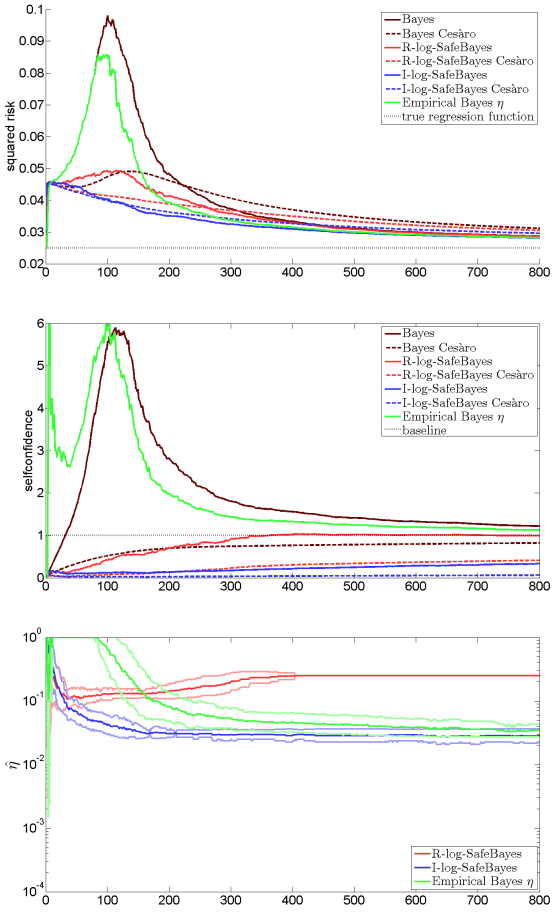

Figure 3.7: Bayesian ridge regression: Model-wrong experiment conditioned on $p := p_{\max} = 50$. The graphs (square-risk, self-confidence ratio and chosen η as function of sample size) are as in Figures 3.3–3.6, except for the third graph there (MAP model order), which has no meaning here. The meaning of the curves is given in Section 3.5.2 except for *empirical Bayes*, explained in Section 3.5.4.

ment, Figure 3.3: when the model is wrong, standard Bayes shows dismal performance in terms of risk and reliability up to a certain sample size and then very slowly recovers, whereas both versions of SafeBayes perform quite well even for small sample sizes. We do not show variations of the graph for $p = p_{\max} = 100$ (i.e. the analogue of Figure 3.5), since it relates to Figure 3.7 in exactly the same way as Figure 3.5 relates to Figure 3.3: with $p = 100$, bad square-risk and reliability behaviour of Bayes goes on for much longer (recovery takes place at much larger sample size) and remains equally good as for

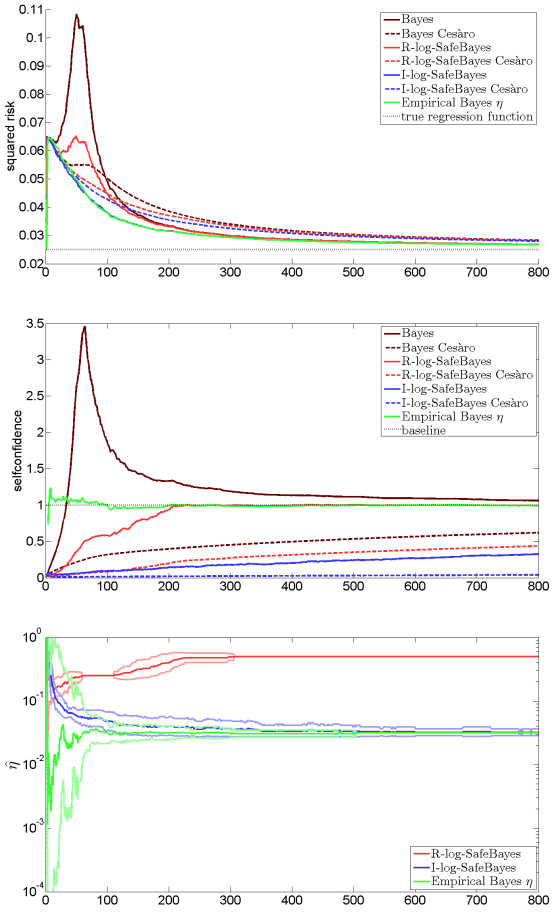

Figure 3.8: Bayesian ridge regression: Same graphs as in Figure 3.7, but for the model-correct experiment conditioned on $p := p_{\max} = 50$.

$p = 50$ with the two versions of SafeBayes.

The results for the Cesàro-versions of our methods are exactly as discussed at the end of Section 3.5.3.

We also see that, as we already indicated in the introduction, choosing the learning rate by empirical Bayes (thus implementing one version of Bayesian bridge regression) behaves terribly. This complies with our general theme that, to 'save Bayes' in general misspecification problems, the parameter η cannot be chosen in a standard Bayesian manner.

Conclusions for model-correct experiment The model-correct experiment for ridge regression (Figure 3.8) offers a surprise: we had expected Bayes to perform best, and were surprised to find that the SafeBayeses obtained smaller

risk. Some followup experiments (not shown here), with different true regression functions and different priors, shed more light on the situation. Consider the setting in which the coefficients of the true function are drawn randomly according to the prior. In this setting standard Bayes performs at least as good in expectation as any other method including SafeBayes (the Bayesian posterior now represents exactly what an experimenter might ideally know). SafeBayes (still in this setting) usually chooses $\eta = 1/2$ or $1/4$, and the difference in risks compared to Bayes is small. On the other hand, if the true coefficients are drawn from a distribution with substantially smaller variance than a priori expected by the prior (a factor 1000 in the 'correct'-model experiment of Figure 3.8), then SafeBayes performs much better than Bayes. Here Bayes can no longer necessarily be expected to have the best performance (the model is correct, but the prior is "wrong"), and it is possible that a slightly reduced learning rate gives (significantly) better results. It seems that this situation, where the variance of the true function is much smaller than its prior expectation, is not exceptional: for example, Raftery et al. (1997) suggest choosing the variance of the prior in such a way that a large region of parameter values receives substantial prior mass. Following that suggestion in our experiments already gives a variance that is large enough compared to the true coefficients that SafeBayes performs better than Bayes even if the model is correct.

A joint observation for the model-wrong and model-correct experiments Finally we note that we see an interesting difference between the two SafeBayes versions here: *I*-log-SafeBayes seems better for risk, giving a smooth decreasing curve in both experiments. *R*-log-SafeBayes inherits a trace of standard Bayes' bad behaviour in both experiments, with a nonmonotonicity in the learning curve. On the other hand, in terms of reliability, *R*-log-SafeBayes is consistently better than *I*-log-SafeBayes (but note that the latter is underconfident, which is arguably preferable over being overconfident, as Bayes is). All in all, there is no clear winner between the two methods.

3.5.5 Executive summary: Joint conclusions from main and additional experiments

Standard Bayes In almost all our experiments (both here and in Chapter 5), standard Bayesian inference fails in its KL-associated prediction tasks (squared error risk, reliability) when the model is wrong. Adopting a different prior (such as the *g*-prior) does not help, with two exceptions in model averaging: (a) when Raftery's prior (Section 5.1.3) is used, then Bayes works quite well, but there it fails dramatically again (in contrast to SafeBayes) once the percentage of easy points is increased; (b) when it is run with a fixed variance that is significantly larger than the 'best' (pseudo-true) variance $\tilde{\sigma}^2$. Moreover, in the ridge regression experiment with fixed σ^2, we find that standard Bayes can even perform much worse than SafeBayes when the model is correct — so all in all we tentatively conclude that SafeBayes is safer to use for linear regression.

SafeBayes R-square-SafeBayes is not competitive with the other SafeBayes methods and can even get worse than Bayes sometimes; this is due to an unwanted dependence on the specified scale σ^2 as explained in Section 5.1. The other three SafeBayes methods behave reasonably well in all our experiments, and there is no clear winner among them. I-square-SafeBayes usually behaves excellently for the square-risk, but cannot directly be used to assess its own performance. I-log-SafeBayes usually behaves excellently in terms of square-risk as well but is underconfident about its own performance (which is perhaps acceptable, overconfidence being a lot more dangerous). R-log-SafeBayes is usually good in terms of square-risk though not as good as I-log-SafeBayes, yet it is highly reliable. However, in Section 5.2.1, we describe an initial idea for discounting the importance of the first few outcomes and explain why this might improve performance. When combined with this discounting idea, R-log-SafeBayes may actually always be competitive with the other two methods in terms of square-risk as well.

Learning η in Bayes- or likelihood way fails Despite its intuitive appeal, fitting η to the data by e.g. empirical Bayes fails both in the model-wrong ridge experiment with a prior on σ^2, where it amounts to Bayesian ridge regression (Figure 3.7) and in the model-wrong fixed-variance ridge experiment (where it amounts to a method for learning the variance, see Section 5.1.1.2).

Robustness of experiments It does not matter whether the $X_{i1}, X_{i2}, \ldots$ are independent Gaussian, uniform or represent polynomial basis functions: all phenomena reported here persist for all choices. If the 'easy' points are not precisely $(0, 0)$, but have themselves a small variance in both dimensions, then all phenomena reported here persist, but on a smaller scale.

Centring We repeated several of our experiments with centred data, i.e. preprocessed data so that the empirical average of the Y_i is exactly 0 (Raftery et al., 1997; Hastie et al., 2001). In none of our experiments did this affect any results. We also looked at the case where the true regression function has an intercept far from 0, and data are *not* centred. This hardly affected the SafeBayes methods.

Other methods We also repeated the wrong-model experiment for several other model selection methods: AIC, BIC, and various forms of cross-validation. Briefly, we found that all these have severe problems with our data as well. experiments, the mentioned methods were used to identify a model index p and η played no role, but in our final experiment we used leave-one-out cross-validation to learn η itself. With the squared error loss it worked fine, which is not too surprising given its close similarity to I-square-SafeBayes. However, when we tried it with log-loss (as a likelihoodist or information-theorist might be tempted to do), it behaved terribly.