Bayesian Analysis (2023) **18**, Number 1, pp. 105–132

A Comparison of Learning Rate Selection Methods in Generalized Bayesian Inference[*]

Pei-Shien Wu[†] and Ryan Martin[‡]

Abstract. Generalized Bayes posterior distributions are formed by putting a fractional power on the likelihood before combining with the prior via Bayes's formula. This fractional power, which is often viewed as a remedy for potential model misspecification bias, is called the *learning rate*, and a number of data-driven learning rate selection methods have been proposed in the recent literature. Each of these proposals has a different focus, a different target they aim to achieve, which makes them difficult to compare. In this paper, we provide a direct head-to-head empirical comparison of these learning rate selection methods in various misspecified model scenarios, in terms of several relevant metrics, in particular, coverage probability of the generalized Bayes credible regions. In some examples all the methods perform well, while in others the misspecification is too severe to be overcome, but we find that the so-called generalized posterior calibration algorithm tends to outperform the others in terms of credible region coverage probability.

MSC2020 subject classifications: Primary 60K35, 60K35; secondary 60K35.

Keywords: coverage probability, generalized posterior calibration algorithm, model misspecification, SafeBayes algorithm.

1 Introduction

Specification of a sound model is a critical part of an effective statistical analysis. This is especially true for a Bayesian approach, since the statistical model or likelihood is explicitly used to construct the posterior distribution from which inferences will be drawn. However, it is common in applications to know relatively little about the phenomenon under investigation, which impacts our ability to specify a sound statistical model. For this reason, the effects of model misspecification have received considerable attention; in the Bayesian literature, this includes Berk (1966), Bunke and Milhaud (1998), Diaconis and Freedman (1986b,a), Kleijn and van der Vaart (2006, 2012), Walker (2013), De Blasi and Walker (2013), Ramamoorthi et al. (2015), and Grünwald and van Ommen (2017). In the most general case, misspecification implies that there is no "true" parameter value that the posterior could concentrate around. Instead, under suitable conditions, the posterior will concentrate around a "best" parameter value, one that minimizes the Kullback–Leibler divergence of the posited model from the true data-generating distribution. But even in those relatively nice cases, where the best parameter value around which the posterior concentrates could be meaningful, or even equal to the real quantity

arXiv: 2012.11349

[*] This work is partially supported by the U.S. National Science Foundation, DMS–1811802.

[†] Department of Statistics, North Carolina State University, pwu9@ncsu.edu

[‡] Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

https://doi.org/10.1214/21-BA1302

of interest, there is reason for concern. Kleijn and van der Vaart (2012) showed that misspecification can also affect the posterior spread, which means that the actual frequentist coverage probability of the Bayesian posterior credible region can be arbitrarily far below the advertised/nominal level.

Real coverage probabilities differing significantly from advertised levels is a serious concern (Fraser, 2011; Martin, 2019). A gap between real and advertised coverage probabilities can have various causes, but here we focus on model misspecification. To avoid this misspecification bias, there are a few options: first, to take an approach that does not depend explicitly on a statistical model; second, to work with a model that is sufficiently broad that misspecification is virtually impossible; and third, to make some adjustments to correct for potential model misspecification. From a Bayesian perspective, the first fix is not available, since Bayes's formula requires a likelihood function. The second fix amounts to the use of Bayesian nonparametrics but, when the quantity of interest is a low-dimensional feature of the full distribution, introducing an infinite-dimensional parameter, with the computational and statistical challenges that entails, would be overkill. This leaves only the third option, but what kind of adjustments might the Bayesian consider? Recently, Grünwald and van Ommen (2017) argued that a certain adjustment to the usual Bayesian posterior distribution can repair inconsistencies resulting from model misspecification. The goal of the present paper is to investigate the extent to which Grünwald and Van Ommen's adjustment—and other related adjustments in the literature—can close the gap between the real and advertised coverage probabilities of Bayesian posterior credible regions affected by model misspecification.

More specifically, here we will be working in the so-called *generalized Bayes* framework, which differs from the traditional Bayes framework only in that a *learning rate* parameter, a power $\eta > 0$ on the likelihood function, is introduced. That is, if we have data D^n and a posited statistical model P_θ^n, indexed by a parameter θ in Θ, then the generalized Bayes posterior distribution for θ is

$$\Pi_n^{(\eta)}(d\theta) \propto L_n^\eta(\theta)\, \Pi(d\theta), \quad \theta \in \Theta, \tag{1.1}$$

where $\theta \mapsto L_n(\theta) = L(\theta; D^n)$ is the likelihood function and Π is a prior distribution on Θ. Among the first papers to adopt such an approach is Walker and Hjort (2001), followed up on by Zhang (2006). Bissiri et al. (2016) showed that (1.1) is the principled way to update prior beliefs when the model is potentially misspecified, and that the appearance of a non-trivial learning rate is a necessary by-product. A different connection between robustness and learning rate $\eta < 1$ was made recently in Miller and Dunson (2019). More details about model misspecification and generalized Bayes are given in Section 2.

Grünwald and Van Ommen's claim is that, for a sufficiently small learning rate η, certain model misspecification biases can be repaired. Of course, the threshold defining "sufficiently small" cannot be known in practice, so some data-driven choices are required. Grünwald (2012, 2018), Grünwald and van Ommen (2017), and de Heide et al. (2020) developed a so-called *SafeBayes* algorithm to choose the learning rate η, based on minimizing a sequential risk measure. A number of other learning rate selection methods have been proposed recently, including the two distinct information matching strategies in Holmes and Walker (2017) and Lyddon et al. (2019), and the bootstrap-motivated calibration method of Syring and Martin (2019). Since the role played by the learning rate is relatively unfamiliar and since the various methods differ significantly in terms of their motivations and implementations, it would be beneficial to see a head-to-head comparison in terms of some standard metrics, for example, the validity and efficiency of the corresponding generalized Bayes credible regions. This paper does just that.

The remainder of this paper is organized as follows. In Section 2, we discuss the behavior of Bayesian posterior distribution under a misspecified model and define and review the literature on generalized Bayes posteriors. For the latter, the choice of learning rate is essential, and we provide details for four recently proposed learning rate selection methods in Section 3. Then, in Section 4, we show a simple illustrative example to give some intuition about how the different methods perform and, in particular, this suggests that the methods which are not designed specifically to calibrate the credible region's coverage probability may not be able to achieve the nominal level in general. In particular, while SafeBayes has been shown to repair inconsistencies, the improved performance is with respect to what Grünwald and van Ommen (2017) call "KL-associated prediction tasks," e.g., achieving small mean-square error in regression, which, of course, provides no guarantees of similarly good performance with respect to inference-related tasks, e.g., credible regions for model parameters achieving the nominal frequentist coverage. To investigate this further, simulation results are presented in Sections 5 and 6 for linear and binary regression models, respectively, and the take-away message is that, while all the learning rate selection methods considered here perform similarly in terms of parameter estimation, when it comes to inference-related tasks, only the inference-focused method of Syring and Martin (2019) can reliably achieve the advertised coverage probability across different sample sizes and misspecification degrees. Some concluding remarks are given in Section 7.

2 Background

2.1 Model misspecification

Suppose we have data D^n which, for simplicity, we assume consists of independent and identically distributed observations: either response variables Y_i only or predictor and response variables pairs (X_i, Y_i), $i = 1, \ldots, n$. Then we posit a statistical model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, a collection of probability measures on the sample space, indexed by a parameter θ taking values in the parameter space Θ. From this model and the observed D^n, we obtain a likelihood function L_n. The likelihood summarizes the information in the data relative to the posited model, which can be combined with prior information encoded in a distribution Π for θ on Θ via Bayes's formula:

$$\Pi_n(d\theta) \propto L_n(\theta)\,\Pi(d\theta), \quad \theta \in \Theta.$$

In the Bayesian paradigm, inferences about θ are drawn based on the posterior distribution Π_n, e.g., degrees of belief about the truthfulness of an assertion "$\theta \in A$," for $A \subset \Theta$, are summarized by the posterior probability $\Pi_n(A)$.

Let $P^\star$ denote the true distribution of Y_1 or of (X_1, Y_1). If the model is correctly specified, then there exists a $\theta^\star \in \Theta$ such that $P^\star = P_{\theta^\star}$. In that case, under suitable regularity conditions, inference based on the posterior distribution will be valid, at least asymptotically. That is, Π_n will concentrate its mass around $\theta^\star$ as $n \to \infty$ and, moreover, the Bernstein–von Mises theorem (e.g., van der Vaart, 2000, Chapter 10) states that Π_n is approximately a normal distribution, centered at the maximum likelihood estimator $\hat{\theta}_n$, with covariance matrix proportional to the inverse of the Fisher information matrix at $\theta^\star$. This implies, among other things, that credible regions derived from Π_n closely resemble those asymptotic confidence regions based on likelihood theory. Therefore, asymptotically, the Bayesian posterior credible regions will have frequentist coverage probability close to the advertised level.

If the model is incorrectly specified, in the sense that $P^\star \notin \mathscr{P}$, then there are several challenges. First, there is no "true" $\theta^\star$, which creates some challenges in interpretation. Indeed, the maximum likelihood estimator, Bayes posterior, or any other model-based procedure will identify the *Kullback–Leibler projection* of $P^\star$ onto the model, i.e.,

$$\theta^\dagger = \arg\min_\theta K(P^\star, P_\theta),$$

where $K(P^\star, P_\theta) = \int \log(dP^\star/dP_\theta)\, dP^\star$ is the Kullback–Leibler divergence of P_θ from $P^\star$. In general, $\theta^\dagger$ does not have a real-world interpretation but, in some cases, certain features of $P^\star$ can be identified based on a misspecified model. For example, if $\mathscr{P}$ is an exponential family, then the mean function of the exponential family model, evaluated at $\theta^\dagger$, equals the mean of $P^\star$ (Bunke and Milhaud, 1998, Example 2). Another similar case is considered in Section 5. The second challenge is that, even in the case where $\theta^\dagger$ has a (limited) real-world interpretation, misspecification can still negatively impact posterior inferences. Kleijn and van der Vaart (2012) established a Bernstein–von Mises theorem under model misspecification which states that, under certain regularity conditions, the posterior Π_n will be approximately normal, with mean equal to the maximum likelihood (or M-) estimator $\hat{\theta}$ and covariance matrix $V_{\theta^\dagger}^{-1}$, where

$$V_{\theta^\dagger} = \int \Bigl(\frac{\partial^2 \log p_\theta}{\partial\theta\partial\theta^\top}\Big|_{\theta=\theta^\dagger} \Bigr)\, dP^\star,$$

and p_θ is the density function corresponding to P_θ. The problem, of course, is that $V_{\theta^\dagger}^{-1}$ is *not* the asymptotic covariance matrix of $\hat{\theta}_n$; the latter, as shown by Huber (1967) and van der Vaart (2000), has the famous sandwich matrix $V_{\theta^\dagger}^{-1}\Lambda_{\theta^\dagger}V_{\theta^\dagger}^{-1}$, where

$$\Lambda_{\theta^\dagger} = \int \Bigl(\frac{\partial \log p_\theta}{\partial\theta}\Big|_{\theta=\theta^\dagger} \Bigr) \Bigl(\frac{\partial \log p_\theta}{\partial\theta}\Big|_{\theta=\theta^\dagger} \Bigr)^\top dP^\star.$$

The implication of this covariance mismatch is that, even if the quantity of interest can be identified under the misspecified model, the frequentist coverage probability of the Bayes posterior credible sets could be arbitrarily far from the advertised level. The question is: *can something be done to correct this problematic behavior?*

2.2 Generalized Bayes

Modifying the usual Bayesian update with a learning rate η as in (1.1) is a simple change, but it has some unexpected consequences. In particular, Walker and Hjort (2001) showed that, for a correctly specified model, consistency of the generalized Bayes posterior $\Pi_n^{(\eta)}$ in (1.1) could be established for any $\eta < 1$, with only local conditions on the prior—as opposed to the local and global conditions required for consistency with $\eta = 1$ (e.g., Ghosal et al., 1999; Barron et al., 1999). The intuition given by Walker et al. (2005) is that inconsistencies result from the posterior over-fitting or tracking the data too closely, and the fractional power discounts the data slightly to prevent this over-fitting. The Walker–Hjort result has been extended to cover posterior concentration rates, where the removal of the global prior conditions—usually formulated in terms of metric entropy (cf. Ghosal et al., 2000; Ghosal and van der Vaart, 2017)—leads to simpler proofs and generally (at least slightly) faster rates. See Zhang (2006) for one of the first papers exploring these ideas, and Bhattacharya et al. (2019) and Grünwald and Mehta (2020) for more recent contributions. The fractional power has also been employed recently in work on high-dimensional problems using an empirical or data-driven prior (e.g., Martin and Walker, 2019; Martin et al., 2017; Martin and Tang, 2020) where, again, the fractional power is motivated by the desire to prevent over-fitting; see, also, Martin (2017) and Martin and Ning (2020) for some potential benefits of $\eta < 1$ to uncertainty quantification.

When the model is misspecified, however, the learning rate is less about convenience and more about necessity. Bissiri et al. (2016) showed that the generalized Bayes update (1.1) is fundamental from a decision-theoretic point of view. Moreover, they argue that the learning rate η naturally emerges since, roughly, the parameter $\theta^\dagger$ being estimated is defined by minimizing the expectation of a loss function $\theta \mapsto \log p_\theta$, and since that minimization problem is invariant to scalar multiples of the loss, the learning rate should appear in the posterior (1.1). In fact, the loss function interpretation makes their result much more general. In many cases, it is more natural to formulate the inference problem with a loss function rather than a statistical model. These are often referred to as *Gibbs posterior distributions*; see Syring and Martin (2017, 2019, 2020b,a), Bhattacharya and Martin (2022), Wang and Martin (2020), and Section 6 below.

Beyond recognizing the importance of the learning rate parameter, an actual value for η needs to be set in practical applications. Several recent papers—including Grünwald and van Ommen (2017), Holmes and Walker (2017), Lyddon et al. (2019), and Syring and Martin (2019)—have proposed data-driven choices for the learning rate, with different motivations. Section 3 describes these methods. The remainder of the paper is focused on a comparison of these different learning rate methods.

3 Learning rate selection methods

3.1 Grünwald's SafeBayes

Grünwald and van Ommen (2017) observe that, when the model is non-convex and misspecified, there is a chance for *hyper-compression*. This is the term they use to describe the seemingly paradoxical result that the Bayesian predictive distribution can be closer, in a Kullback–Leibler sense, to the true $P^\star$ than the within-model Kullback–Leibler minimizer $P_{\theta^\dagger}$. What makes this possible is non-convexity: the predictive distribution is an average of in-model distributions P_θ which, without convexity, could end up outside the model and potentially closer to $P^\star$ than is $P_{\theta^\dagger}$. Besides being counter-intuitive, hyper-compression also reveals a practical problem, namely, inconsistency—that the posterior distribution is not concentrating its mass near $\theta^\dagger$ as expected. To overcome this, Grünwald and van Ommen (2017) suggest to work with a new (hypothetical) model, with densities

$$p_\theta^{(\eta)}(x,y) = p^\star(x,y)\bigl\{p_\theta(y \mid x)/p_{\theta^\dagger}(y \mid x)\bigr\}^\eta,$$

indexed by a parameter $\eta > 0$. We say this model is "hypothetical" because it depends on $p^\star$ and $\theta^\dagger$, two ingredients that are not available to the data analyst. However, *if η is sufficiently small*, in the sense that $\int p_\theta^{(\eta)}(x,y)\,dx\,dy$ is strictly less than 1, then this indeed defines a genuine statistical model, with two interesting properties:

- it is not misspecified, i.e., the Kullback–Leibler minimizer is $\theta^\dagger$ and $p_{\theta^\dagger}^{(\eta)} = p^\star$;
- and the Bayesian posterior based on this new model is precisely the generalized Bayes posterior $\Pi_n^{(\eta)}$, with learning rate η, as in (1.1).

Since this new model is not misspecified, hyper-compression and inconsistency of the η-generalized Bayes posterior can be avoided. So: *how to choose η sufficiently small?*

Grünwald and van Ommen (2017), building on work in, e.g., Grünwald (2012), argue that the so-called *SafeBayes* algorithm will select a learning rate η that is sufficiently small in the sense above. Define the cumulative expected log-loss under the η-generalized Bayes posterior distribution, as a function of η:

$$\eta \mapsto \sum_{i=1}^n \int -\log p_\theta(Y_i \mid X_i)\,\Pi_{i-1}^{(\eta)}(d\theta). \tag{3.1}$$

The SafeBayes algorithm returns the minimizer, $\hat\eta$, of this function over the range $\eta \in [0,1]$. Grünwald (2012) presents an argument for why the SafeBayes choice of $\hat\eta$ works in the sense of being sufficiently small as in the discussion above. Note that it is SafeBayes's focus on minimizing the cumulative log-loss that makes it especially suited for overcoming misspecification with respect to "KL-associated prediction tasks."

What we have described here is one of two versions of the SafeBayes algorithm presented in Grünwald and van Ommen (2017), namely, the "R-SafeBayes" version. In our examples below, we found that the "R" version outperformed the other—namely, the "I-SafeBayes" version—so here we only discuss the former.

3.2 Holmes and Walker (2017)

Following Bissiri et al. (2016), the Bayesian and generalized Bayesian frameworks can be considered simply as rules for using data to update prior beliefs to posterior beliefs. As such, it makes sense to consider how much information has been gained from the update, by comparing the prior to the posterior. Of course, this information gain depends on both the updating rule and on the data, and Holmes and Walker (2017) proposed a procedure for selecting the learning rate η based on matching the expected information gain between Bayes and generalized Bayes updates.

More formally, if $I_\eta(x,y)$ denotes the information gain in the generalized Bayes update from prior to posterior based on learning rate η and data values (x,y), then Holmes and Walker (2017) propose to set η such that

$$\int I_\eta(x,y)\, P^\star(dx,dy) = \int I_1(x,y)\, P_{\theta^\dagger}(dx,dy), \tag{3.2}$$

where $I_1(\cdot)$ denotes the information gain in the standard Bayesian update. The specific choice of information measure they recommend is the *Fisher divergence*

$$I_\eta(x,y) = \int \{\nabla \log \pi_{x,y}^{(\eta)}(\theta) - \nabla \log \pi(\theta)\}^2\, \pi(\theta)\, d\theta,$$

where $\pi_{x,y}^{(\eta)}$ denotes the generalized Bayes posterior based on data (x,y) and learning rate η, and ∇ is the gradient operator. Then it is straightforward to check that $I_\eta(x,y) = \eta^2 I_1(x,y)$ and, therefore, by (3.2), an "oracle" learning rate is given by

$$\eta^\star = \Big\{ \frac{\int I_1(x,y)\, P_{\theta^\dagger}(dx,dy)}{\int I_1(x,y)\, P^\star(dx,dy)} \Big\}^{1/2}.$$

Of course, both $P^\star$ and $P_{\theta^\dagger}$ are unknown, so $\eta^\star$ cannot be evaluated, but the expectations can be estimated with the actual data $\{(X_i,Y_i) : i = 1,\ldots,n\}$. That is,

$$\hat\eta = \Big\{ \frac{\int I_1(x,y)\, P_{\hat\theta_n}(dx,dy)}{\int I_1(x,y)\, \mathbb{P}_n(dx,dy)} \Big\}^{1/2},$$

where $\hat\theta_n$ is the maximum likelihood estimator of θ under the model—which is an estimate of $\theta^\dagger$—and $\mathbb{P}_n$ is the empirical distribution of the data.

3.3 Lyddon et al. (2019)

The learning rate selection strategy presented in Lyddon et al. (2019) is motivated by the weighted likelihood bootstrap approach of Newton and Raftery (1994), which was shown to generate bootstrap samples that have the same asymptotic distribution as Bayesian posterior distribution under a correctly specified model. For the case of a misspecified model, Lyddon et al. (2019) proposed a modified weighted likelihood bootstrap approach which replaces the ordinary bootstrap with the Bayesian bootstrap, and establish its asymptotic limiting distribution. Then following a strategy similar to that in Holmes and Walker (2017) described above, they propose to choose η in order to match the limiting η-generalized Bayes posterior to that of this modified likelihood bootstrap. They then show that, using the notation defined at the end of Section 2.1, an "oracle" learning rate is

$$\eta^\star = \frac{\operatorname{tr}(V_{\theta^\dagger} \Lambda_{\theta^\dagger}^{-1} V_{\theta^\dagger})}{\operatorname{tr}(V_{\theta^\dagger})}.$$

Again, since $\theta^\dagger$ and $P^\star$ are unknown, this oracle value cannot be evaluated. However, a data-driven choice $\hat{\eta}$ can be obtained by replacing $\theta^\dagger$ the maximum likelihood estimator and the expectations with respect to $P^\star$ in V_θ and Λ_θ, respectively, with expectations with respect to the empirical distribution $\mathbb{P}_n$.

3.4 Syring and Martin (2019)

The three previous subsections describe principled learning rate selection strategies, but none of those are tailored so that the generalized posterior distribution achieves any specific and desirable frequentist properties. Since the learning rate's effect on the posterior is largely to control the spread, Syring and Martin (2019) proposed to tune the learning rate such that posterior credible sets approximately achieve the nominal frequentist coverage probability.

The coverage probability function is given by

$$c_\alpha(\eta \mid P^\star) = P^\star\{C_\alpha^{(\eta)}(D^n) \ni \theta^\dagger\},$$

where $C_\alpha^{(\eta)}$ is the η-generalized Bayes $100(1-\alpha)\%$ credible region for θ, e.g., a highest posterior density region, $\theta^\dagger$ is the Kullback–Leibler minimizer in the model, treated as a functional of $P^\star$, and $D^n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ is the iid data set from $P^\star$. Then the goal is to find η such that $c_\alpha(\eta \mid P^\star) = 1 - \alpha$. Of course, lots of the quantities involved in this equation are unknown, but they can be estimated. In particular, if $P^\star$ is replaced by the empirical distribution $\mathbb{P}_n$, then the new equation is

$$c_\alpha(\eta \mid \mathbb{P}_n) := \mathbb{P}_n\{C_\alpha^{(\eta)}(D^n) \ni \hat{\theta}_n\} = 1 - \alpha,$$

where $\hat{\theta}_n$ is the maximum likelihood estimator based on the observed data, i.e., the "$\theta^\dagger$-functional" applied to $\mathbb{P}_n$. Even this alternative coverage probability function cannot be evaluated, since it requires enumerating all n^n possible with-replacement samples from the observed data, but a bootstrap approximation is possible. That is, for B bootstrap samples $\tilde{D}_1^n, \ldots, \tilde{D}_B^n$, calculate

$$\hat{c}_\alpha(\eta \mid \mathbb{P}_n) = \frac{1}{B} \sum_{b=1}^{B} 1\{C_\alpha^{(\eta)}(\tilde{D}_b^n) \ni \hat{\theta}_n\}.$$

To solve the equation, $\hat{c}_\alpha(\eta \mid \mathbb{P}_n) = 1 - \alpha$, Syring and Martin (2019) recommend a stochastic approximation scheme that defines a learning rate sequence (η_t) as

$$\eta_{t+1} = \eta_t + k_t\{\hat{c}_\alpha(\eta \mid \mathbb{P}_n) - (1 - \alpha)\}, \quad t \geq 1,$$

where k_t is a sequence such that $\sum_t k_t = \infty$ and $\sum_t k_t^2 < \infty$. When the η_t sequence effectively converges, the limit is the suggested learning rate $\hat{\eta}$. This is what Syring and Martin (2019) refer to as the generalized posterior calibration (GPC) algorithm. Like SafeBayes, it is relatively expensive computationally—these algorithms require posterior computations for multiple learning rates and data sets—but the benefit is having a posterior distribution with meaningful spread, even in finite samples.

4 Learning rates in a toy example

Before we consider the effect of different learning rate selection methods in some nontrivial real-world problems, it helps to consider a simple example, one where some of the calculations can be done by hand, to develop some intuition about what to expect.

Suppose that the posited model for iid data $Y^n = (Y_1, \ldots, Y_n)$ is $P_\theta = \mathsf{N}(\theta, \sigma^2)$, with $\sigma > 0$ fixed, but that the true distribution is $P^\star = \mathsf{N}(\theta^\star, \sigma^{\star 2})$, where $\sigma^\star > 0$ is potentially different from σ. With a conjugate normal prior, $\theta \sim \mathsf{N}(0, \sigma^2/\sigma_0^2)$ where $\sigma_0^2 = 10$, the η-generalized Bayes posterior density is $\Pi_n = \mathsf{N}(m_n, v_n)$, with

$$m_n = \frac{n\eta}{n\eta + \sigma_0^2}\bar{y}_n \quad \text{and} \quad v_n = \frac{\sigma^2}{n\eta + \sigma_0^2}.$$

It is easy to confirm that the Kullback–Leibler minimizer satisfies $\theta^\dagger = \theta^\star$, but the misspecified variance can still cause problems, as we now demonstrate.

It is intuitively clear that the generalized Bayes framework could completely resolve the model misspecification if the learning rate was chosen as $\eta^\star = (\sigma/\sigma^\star)^2$. More formally, de Heide et al. (2020) show that, if the learning rate is no larger than this ratio, the generalized Bayes posterior will enjoy fast root-n rate convergence properties, while Syring and Martin (2019) argue that, in this and other similar problems, taking the learning rate exactly equal to $\eta^\star$ is necessary in order to achieve exact coverage of credible sets. In any case, of course, one cannot make this learning rate choice in practice because it depends on the unknown value of true variance. But this intuition tells us what the different learning rate selection methods' target should be.

To evaluate the performance of the different learning rate methods, we simulate 1000 data sets, for each of several different sample sizes n and values of $\eta^\star = (\sigma/\sigma^\star)^2$, and compare the average estimated learning rate against $\eta^\star$; see Figure 1. If the estimated η is close to the diagonal line $\eta = \eta^\star$, then the generalized Bayesian credible sets have coverage probability near the nominal level. To confirm this, see Figure 2. When the degree of misspecification is relatively mild, i.e., when $\eta^\star \approx 1$, all the methods perform well. However, as the misspecification degree increases, or $\eta^\star$ decreases, we find that SafeBayes and the Holmes and Walker method have decreasing coverage probability, quickly falling below any reasonable tolerance. On the other hand, both the Lyddon et al. and Syring and Martin methods are able to achieve the target 95% coverage probability over the entire range of settings.

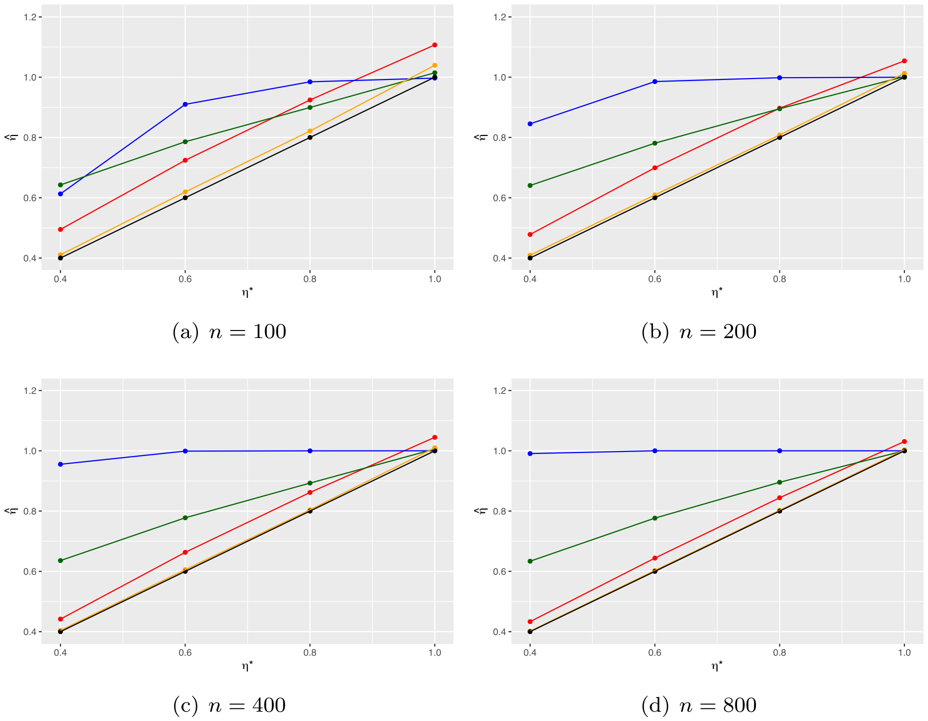
5 Learning rates in linear regression

5.1 Model setup

Consider a linear regression model of the form

$$y_i = x_i^\top \beta + \sigma\varepsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

where the pairs $(x_1, y_1), \ldots, (x_n, y_n)$, taking values in $\mathbb{R}^p \times \mathbb{R}$, are independent, $\beta \in \mathbb{R}^p$ is an unknown vector of coefficients, σ is an unknown scale parameter, and $\varepsilon_1, \ldots, \varepsilon_n$



(a) $n = 100$

(b) $n = 200$

(c) $n = 400$

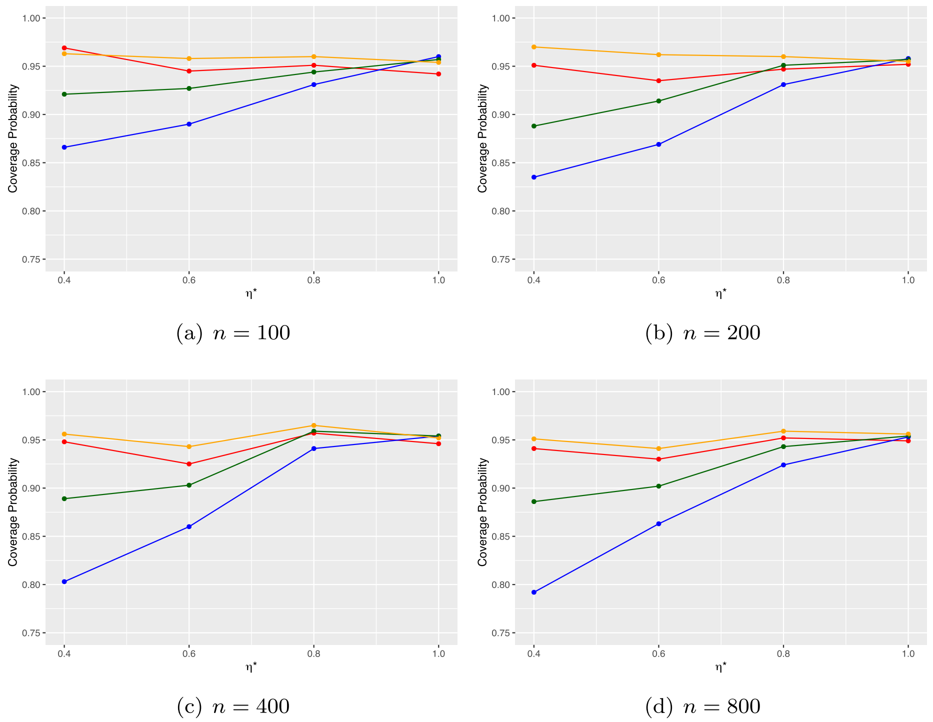
(d) $n = 800$

Figure 1: Average learning rate η, across 1000 replications, versus the optimal $\eta^\star = (\sigma/\sigma^\star)^2$, the closer to the diagonal line the better. True (black), GPC (red), R-SafeBayes (blue), Holmes and Walker (green), Lyddon et al. (orange).

are random error terms. As is most common, here we will consider a model that assumes the errors $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathsf{N}(0,1)$, independent of $x_1, \ldots, x_n$. In the experiments that follow, the p-dimensional covariates, x_i, are taken to be iid from a multivariate normal distribution with mean zero, unit variance, and a first-order autocorrelation structure, i.e., $\mathsf{E}(x_{ij}x_{ik}) = \rho^{|j-k|}$, with correlation $\rho = 0.2$. For most of this section, we take $p = 4$ and set the true coefficient vector to be $\beta = (1, 1, 2, -1)^\top$.

If it happens that the true distribution is different from this posited model, then, in general, we can expect an ordinary Bayes posterior to suffer from misspecification bias. The goal here is to investigate how the different learning rate methods can help the generalized Bayes posterior to correct for this misspecification bias.

The two key assumptions behind the textbook linear regression model are that the errors are (a) independent of covariates and (b) normally distributed. Here we present results for two types of misspecification, namely, *Dependent Errors* and *Non-normal Errors*. The specific form and degree of these misspecifications will be described in the following subsections. For the comparison, the metrics we consider are



(a) $n = 100$

(b) $n = 200$

(c) $n = 400$

(d) $n = 800$

Figure 2: Average coverage probability of the nominal 95% generalized Bayes credible intervals, across 1000 replications. GPC (red), SafeBayes (blue), Holmes and Walker (green), Lyddon et al. (orange).

- mean value of the learning rate estimates, $\hat{\eta}$;
- coverage probability of the $\hat{\eta}$-generalized Bayes 95% highest posterior density credible sets for the full β vector;
- mean square error of the $\hat{\eta}$-generalized Bayes posterior mean of β;
- the average of the marginal $\hat{\eta}$-generalized Bayes posterior variances for each coordinate of β.

We are specifically interested in the learning rate and its effect on the coverage probability of the generalized posterior credible sets, so the first two metrics are clear. The mean square error of the generalized posterior mean acts like an overall measure of bias, i.e., how far does the center of the posterior tend to be from the true parameter values. In the examples that follow, we find that the mean square error does not vary much relative to the learning rate selection method, which confirms our intuition that, at least within a suitable range of η values, the choice of learning rate really only impacts the posterior spread; as we show in Section 5.4, this remains true even as the dimension of the covariate increases. The fourth metric is an overall measure of the spread of the generalized Bayes posterior, and we expect that those learning rate selection methods whose credible regions tend to under-cover will have smaller total variance.

Let $\mathsf{Gamma}(a, b)$ denote a gamma distribution with shape parameter $a > 0$ and rate parameter $b > 0$; the density function is

$$y \mapsto y^{a-1}e^{-by}, \quad y > 0.$$

For a Bayesian analysis, we proceed by introducing a conjugate normal–inverse gamma prior for (β, σ^2), where the conditional prior for β, given σ^2, is $\mathsf{N}_p(0, \sigma^2 I_p)$ and the marginal prior for σ^{-2} is $\mathsf{Gamma}(a = 1, b = 0.025)$. Then the η-generalized posterior is

$$\begin{aligned}(\sigma^{-2} \mid y, X, \eta) &\sim \mathsf{Gamma}\left(a = a_0 + \tfrac{1}{2}\eta n,\, b = b_0 + \tfrac{\eta}{2}(y^\top y - m_n^\top X^\top y)\right)\\ (\beta \mid \sigma^2, y, X, \eta) &\sim \mathsf{N}_p(m_n, \sigma^2 V_n),\end{aligned}$$

where $V_n = (\eta X^\top X + I)^{-1}$ and $m_n = \eta V_n X^\top y$.

The method of Holmes and Walker and Lyddon et al. are much less demanding in terms of computation time compared to GPC and SafeBayes, so here we only compare the computational time of the latter two and only in the highest degree of misspecification cases. For the SafeBayes algorithm, we follow the recommendation of Grünwald and van Ommen (2017) and set the grid points for η as $\{1, 2^{-1/3}, 2^{-2/3}, \ldots, 2^{-8}\}$ and calculate the total computation time. For the GPC algorithm, we calculate the total computation time that the credible regions of the generalized posterior distribution had empirical coverage probability near the nominal level, based on 200 bootstrap samples. Both GPC and SafeBayes take less than 7 seconds per replicate, even for $n = 400$, though SafeBayes is faster than GPC in this case. This is because a closed-form expression for the right-hand side of (3.1) is available in this example; GPC, on the other hand, can use a similar closed-form expression, but it still has to cycle through each bootstrap sample. The computation time comparisons change dramatically, however, once we move beyond the simple conjugate linear regression model; see Section 6.

5.2 Misspecified error distribution

Dependent errors

Here we deviate from the textbook linear model assumptions by allowing the variance of the error ε_i to depend in a certain way on the covariate vector x_i, $i = 1, \ldots, n$. In particular, let $\hat{\xi}_{0.05}$ and $\hat{\xi}_{0.95}$ denote the sample 5th and 95th percentiles of $x_{11}, \ldots, x_{n1}$. Then define the case-specific standard deviation as

$$\sigma_i = \begin{cases} s_{\text{small}} & \text{if } x_{i1} < \hat{\xi}_{0.05} \\ s_{\text{mod}} & \text{if } \hat{\xi}_{0.05} \le x_{i1} \le \hat{\xi}_{0.95} \\ 1 & \text{if } x_{i1} > \hat{\xi}_{0.95}, \end{cases}$$

where the small and moderate values, s_{small} and s_{mod}, control the degree of the departures from constant variance. We consider three different degrees of misspecification.

Degree 1. $s_{\text{small}} = 0.25$ and $s_{\text{mod}} = 0.50$;

Degree 2. $s_{\text{small}} = 0.05$ and $s_{\text{mod}} = 0.25$;

Degree 3. $s_{\text{small}} = 0.01$ and $s_{\text{mod}} = 0.10$.

A summary of the different learning rate selection procedures, across the different misspecification degrees and sample sizes $n \in \{100, 200, 400\}$, is presented in Table 1, based on 1000 data sets for each combination. In Degree 1, where the misspecification is relatively mild, we see that all four learning rate selections perform well and similarly in terms of both the learning rates chosen—all near 1—and in the coverage probabilities. As expected, however, as misspecification gets more severe, in Degrees 2 and 3, the more disparity we see between the selected learning rates and, in turn, in the coverage probabilities. Only GPC is able to achieve the nominal coverage probability in the more severe misspecification settings, while the performance of other methods can be quite poor, especially under Degree 3 with small n. The mean square errors are more or less the same for the methods within each sample size–degree combination; and the fact that these values are small indicates the posterior is generally centered around the true β values. As for the posterior spread, there is not much difference between the results in the Degree 1 case with only mild misspecification. However, in Degrees 2 and 3, where the misspecification is more severe, we see greater difference in the posterior variance. As expected, those methods whose posterior variance tends to be small are those who tend to have credible sets that under-cover, in many cases severely.

Non-normal errors

Next, we consider departures from the specified model in terms of the distribution of the error terms. It turns out that the performance of the learning rate selection methods was less sensitive to departures from normality compared to departures from the constant-error-variance assumption. Here we present the results for only one kind of departure from normality, namely, with heavy-tailed errors. In particular, consider errors $\varepsilon_1, \ldots, \varepsilon_n$ iid from a Student-t distribution with degrees of freedom ν. As before, we consider three degrees of misspecification, each sufficiently light-tailed that the variance exists.

Degree 1. $\nu = 5$;

Degree 2. $\nu = 4$;

Degree 3. $\nu = 3$.

Table 2 summarizes the results just like in the previous subsection. Here, however, the differences in performance across different learning rate selection methods, sample sizes, and misspecification degrees is much smaller. Overall the methods return similar learning rate estimates and hit the target coverage probability on the mark. The method of Lyddon et al. tends to select a learning rate that is too large, leading to under-coverage, but its performance tends to improve as the sample size increases.

Degree	n	Method	$\hat{\eta}$	Coverage	MSE	Variance
1	100	GPC	0.95	0.95	0.05	0.012
		SafeBayes	0.92	0.94	0.05	0.014
		Holmes and Walker	1.00	0.93	0.05	0.011
		Lyddon et al.	1.18	0.89	0.05	0.010
	200	GPC	0.95	0.93	0.02	0.006
		SafeBayes	0.92	0.93	0.02	0.007
		Holmes and Walker	0.99	0.92	0.02	0.006
		Lyddon et al.	1.06	0.90	0.02	0.005
	400	GPC	0.94	0.95	0.01	0.003
		SafeBayes	0.93	0.94	0.01	0.003
		Holmes and Walker	0.99	0.94	0.01	0.003
		Lyddon et al.	0.99	0.94	0.01	0.003
2	100	GPC	0.79	0.95	0.06	0.015
		SafeBayes	0.90	0.90	0.06	0.014
		Holmes and Walker	0.98	0.89	0.06	0.012
		Lyddon et al.	1.33	0.76	0.06	0.009
	200	GPC	0.75	0.95	0.03	0.008
		SafeBayes	0.92	0.90	0.03	0.006
		Holmes and Walker	0.97	0.89	0.03	0.006
		Lyddon et al.	1.11	0.84	0.03	0.005
	400	GPC	0.74	0.94	0.01	0.004
		SafeBayes	0.93	0.89	0.01	0.003
		Holmes and Walker	0.96	0.88	0.01	0.003
		Lyddon et al.	0.97	0.88	0.01	0.003
3	100	GPC	0.54	0.98	0.07	0.023
		SafeBayes	0.75	0.87	0.07	0.018
		Holmes and Walker	0.94	0.80	0.07	0.012
		Lyddon et al.	2.45	0.38	0.07	0.005
	200	GPC	0.53	0.95	0.04	0.011
		SafeBayes	0.76	0.86	0.04	0.008
		Holmes and Walker	0.91	0.79	0.04	0.006
		Lyddon et al.	1.74	0.53	0.04	0.003
	400	GPC	0.53	0.95	0.02	0.005
		SafeBayes	0.78	0.84	0.02	0.004
		Holmes and Walker	0.89	0.81	0.02	0.003
		Lyddon et al.	1.25	0.69	0.02	0.002

Table 1: Comparison of average learning rate estimates ($\hat{\eta}$), estimated coverage probabilities (Coverage), mean square error (MSE), and total posterior variance (Variance) across different sample sizes and misspecification degrees in the *Dependent Errors* example.

Degree	n	Method	$\hat{\eta}$	Coverage	MSE	Variance
1	100	GPC	0.98	0.95	0.07	0.019
		SafeBayes	0.90	0.95	0.07	0.023
		Holmes and Walker	1.00	0.94	0.07	0.019
		Lyddon et al.	1.28	0.87	0.07	0.014
	200	GPC	0.99	0.96	0.04	0.009
		SafeBayes	0.90	0.96	0.04	0.011
		Holmes and Walker	0.98	0.96	0.04	0.009
		Lyddon et al.	1.15	0.92	0.04	0.008
	400	GPC	1.00	0.95	0.02	0.005
		SafeBayes	0.92	0.95	0.02	0.005
		Holmes and Walker	0.98	0.95	0.02	0.005
		Lyddon et al.	1.07	0.94	0.02	0.004
2	100	GPC	0.97	0.96	0.08	0.024
		SafeBayes	0.89	0.96	0.08	0.028
		Holmes and Walker	0.99	0.96	0.08	0.023
		Lyddon et al.	1.38	0.87	0.08	0.016
	200	GPC	0.98	0.96	0.04	0.011
		SafeBayes	0.91	0.96	0.04	0.013
		Holmes and Walker	0.97	0.96	0.04	0.011
		Lyddon et al.	1.20	0.91	0.04	0.009
	400	GPC	0.99	0.96	0.02	0.006
		SafeBayes	0.92	0.96	0.02	0.007
		Holmes and Walker	0.96	0.96	0.02	0.006
		Lyddon et al.	1.12	0.92	0.02	0.005
3	100	GPC	0.92	0.97	0.13	0.044
		SafeBayes	0.87	0.97	0.13	0.047
		Holmes and Walker	0.93	0.96	0.13	0.041
		Lyddon et al.	1.69	0.81	0.13	0.020
	200	GPC	0.96	0.97	0.06	0.018
		SafeBayes	0.90	0.96	0.06	0.021
		Holmes and Walker	0.92	0.96	0.06	0.020
		Lyddon et al.	1.37	0.86	0.06	0.011
	400	GPC	0.97	0.96	0.03	0.009
		SafeBayes	0.91	0.96	0.03	0.010
		Holmes and Walker	0.86	0.96	0.03	0.012
		Lyddon et al.	1.26	0.89	0.03	0.006

Table 2: Comparison of average learning rate estimates ($\hat{\eta}$), estimated coverage probabilities (Coverage), mean square error (MSE), and total posterior variance (Variance) across different sample sizes and misspecification degrees in the *Non-normal Errors* example.

Other experiments

Finally, we considered other types of misspecification in addition to those presented above. These results are not presented here because all four learning rate selection methods performed similarly and displaying a table of similar numbers is not a good use of space. But it is worth mentioning in what cases these methods perform comparably, and below is a brief summary of our findings.

- In cases where the heteroscedasticity is less extreme than in Section 5.2 above, in particular, with errors having non-constant variance but independent of x, we found that all four learning rate selection methods performed well. That is, the learning rate estimates were all similar and the credible regions all had coverage probability near the nominal 95% level.
- The example in Section 5.2 considered heavy-tailed error distributions. We also considered cases where the error distribution was asymmetric, e.g., skew-normal (Pérez-Rodríguez et al., 2018). Apparently, misspecification in the shape of the error distribution has little effect because, as above, the learning rate selection methods all performed well in these cases.

5.3 Real data analysis

To compare different learning rate selection methods in a real data set, we can do resampling from the observed data. Suppose that we have data D^n, a posited statistical model P_θ, and the maximum likelihood estimate is $\hat{\theta}$. For resampling, let $\{D_r^n\}_{r=1}^R$ be the R resamples, each of size n, generated from D^n. That is, each D_r^n is a random sample of size n, with replacement, from the original data D^n. We then estimate the learning rate using different learning rate selection methods on the resampled data, and investigate whether the credible regions for the various $\hat{\eta}$-generalized posteriors can cover $\hat{\theta}$. In particular, the coverage probability is estimated by

$$\frac{1}{R}\sum_{i=1}^{R} 1\{C_{\hat{\eta},\alpha}(D_r^n) \ni \hat{\theta}\},$$

where $C_{\hat{\eta},\alpha}(D_r^n)$ is the $100(1-\alpha)\%$ credible region for θ from the $\hat{\eta}$-generalized posterior based on resampled data D_r^n, $r = 1, \ldots, R$.

Here we consider the data of the number of plant species along with several geographic variables, the *gala* data in R `faraway` package (Faraway, 2016). Figure 3 demonstrates that a textbook normal linear model is severely misspecified in the sense that the errors are clearly heteroscedastic, i.e., have non-constant variance. Among the four learning rate selection methods, the Lyddon et al. method selects a learning rate that is far too large, leading to poor coverage, so here we only compare the performance of the other three learning rate selection methods. Table 3 presents the estimated coverage probabilities for the three learning rate selection methods, based on $R = 1000$ resamples. Clearly, for a severely misspecified model like this, only GPC is able to select a small enough learning rate to achieve the desired coverage.

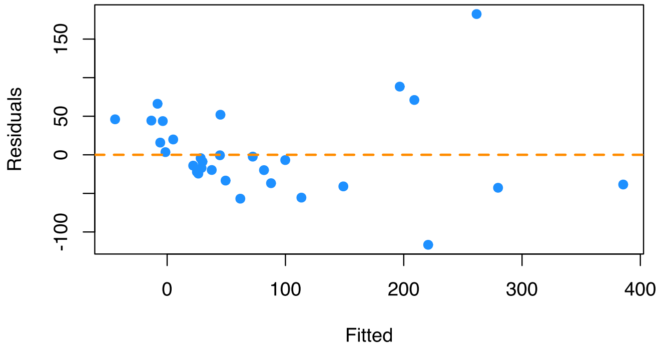


Figure 3: Plot of the residuals versus fitted values based on the real data example in Section 5.3, based on a textbook normal linear regression model.

Method	$\hat{\eta}$	Coverage	MSE	Variance
GPC	0.054	0.963	2.231	11.889
SafeBayes	0.154	0.916	2.363	5.335
Holmes and Walker	1.003	0.477	2.556	0.172

Table 3: Comparison of average learning rate estimates ($\hat{\eta}$), estimated coverage probabilities (Coverage), mean square error (MSE), and total posterior variance (Variance) in the *gala data* example.

5.4 Effect of covariate dimension

Previous experiments focused on how the degree of misspecification affects the behavior of generalized posterior distributions. Instead of varying the misspecification degree, it would be interesting to see how different learning rate selection methods perform across different covariate dimensions $p \in \{4, 8, 16, 32\}$ when the sample size is fixed and relatively small, $n = 100$. We generate the covariates x_i in the same way as in Section 5.2, and extended the low-dimensional linear regression model examples to the more general case. The true coefficient vector recycles the four values $(1, 1, 2, -1)^\top$ used in the previous simulations for each p, which are multiples of 4. To demonstrate the effect of covariate dimension, here we consider a relatively mild misspecification case, degree 2 misspecification in the *dependent errors* example.

The results are summarized in Table 4. First, notice that the learning rate selection procedure does not have much effect on the posterior mean's MSE, which reiterates our previous claim that the learning rate mainly controls the posterior spread, and has little effect on the posterior center. Second, we can see that GPC can still achieve the $100(1-\alpha)\%$ coverage when the dimension is pushed to $p = 16$ at small sample size $n = 100$. On the other hand, SafeBayes and Lyddon et al. tend to choose too small and too large of learning rates, leading to over- and under-coverage, respectively, as the dimension increases. When the dimension is relatively large, all but Lyddon et al. tend to be quite conservative. This is perhaps not surprising, since adjustments being made

p	Method	$\hat{\eta}$	Coverage	MSE	Variance
4	GPC	0.79	0.95	0.06	0.015
	SafeBayes	0.90	0.90	0.06	0.014
	Holmes and Walker	0.98	0.89	0.06	0.012
	Lyddon et al.	1.33	0.76	0.06	0.009
8	GPC	0.94	0.95	0.10	0.013
	SafeBayes	0.80	0.95	0.11	0.019
	Holmes and Walker	1.00	0.93	0.10	0.012
	Lyddon et al.	1.67	0.60	0.10	0.007
16	GPC	0.98	0.96	0.22	0.015
	SafeBayes	0.57	1.00	0.23	0.037
	Holmes and Walker	1.01	0.94	0.22	0.014
	Lyddon et al.	2.20	0.24	0.22	0.006
32	GPC	1.01	0.99	0.52	0.019
	SafeBayes	0.38	1.00	0.63	0.103
	Holmes and Walker	1.01	1.00	0.52	0.019
	Lyddon et al.	3.46	0.00	0.51	0.004

Table 4: Comparison of average learning rate estimates ($\hat{\eta}$), estimated coverage probabilities (Coverage), mean square error (MSE), and total posterior variance (Variance) across different dimensions (p) in the *Dependent Errors* misspecified degree-2 example with $n = 100$.

to shrink/stretch the credible regions are overly simple and do not account for any structure (e.g., sparsity) in the higher-dimensional parameters.

6 Learning rates in logistic regression

6.1 Model setup

An important problem in medical statistics is estimation of the so-called *minimum clinically important difference* (MCID) that assesses the practical as opposed to statistical significance of a treatment. In words, the MCID is the threshold on the diagnostic measure scale such that improvements beyond that level are associated with patients feeling better after the treatment; see, e.g., Hedayat et al. (2015) and the references therein. To set the scene, let $X \in \mathbb{R}$ denote the patient's diagnostic measure, e.g., the pre-treatment minus post-treatment difference in blood pressure, and let $Y \in \{-1, +1\}$ denote the patient-reported indicator of whether they felt the treatment was effective, with "$y = +1$" indicating effective. The quantity of interest, θ, the MCID, is the cutoff on the X scale such that the indicator $1\{X > \theta\}$ is most highly associated with Y. More precisely, the MCID is defined as

$$\theta = \arg\min_{\vartheta} P\{Y \neq \text{sign}(X - \vartheta)\},$$

where $\text{sign}(0) = 1$. Clearly, θ depends on the unknown joint distribution P of (X, Y).

Towards inference on the MCID, it is natural to introduce a statistical model for P. It would be difficult to develop a model for which θ is directly a model parameter, but one idea would be to use a logistic regression model with Y as the binary response and X as a continuous predictor. That is, the logistic regression model states that

$$(Y \mid X = x) \sim \mathsf{Rad}\big(F(\beta_0 + \beta_1 x)\big),$$

where $\mathsf{Rad}(p)$ denotes a Rademacher distribution, i.e., a binary distribution on $\{-1, +1\}$, with probability mass p assigned to the value $+1$, and F is a logistic distribution function with $F(u) = (1 + e^{-u})^{-1}$, for $u \in \mathbb{R}$. The logistic regression model is determined by the unknown parameters (β_0, β_1). Since the MCID θ also satisfies $P(Y = +1 \mid X = \theta) = \frac{1}{2}$, if the above model is assumed, then

$$\theta = -\beta_0/\beta_1.$$

Given independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from this model, a posterior distribution for (β_0, β_1), generalized Bayes or otherwise, can be obtained. From this, one can readily obtain the corresponding posterior distribution of θ via the identity above.

Of course, this model could easily be misspecified. So it is of interest to investigate what happens with the generalized Bayes posterior with suitably chosen learning rates when the logistic link function F is incorrectly specified.

6.2 Results

We fit a misspecified logistic regression model, i.e., where the diagnostic measure X comes from a normal mixture model with distribution function

$$F^\star(x) = 0.7\, \Phi(x \mid 5, 1) + 0.3\, \Phi(x \mid \mu, 1),$$

and the patient reported effectiveness indicator is $(Y \mid X = x) \sim \mathsf{Rad}(F^\star(x))$. The quantity μ controls the degree of misspecification, with μ closer to 5 corresponding to "less misspecification" relative to the logistic link function F above; see Figure 4. The three specific degrees considered are:

Degree 1. $\mu = 7$;

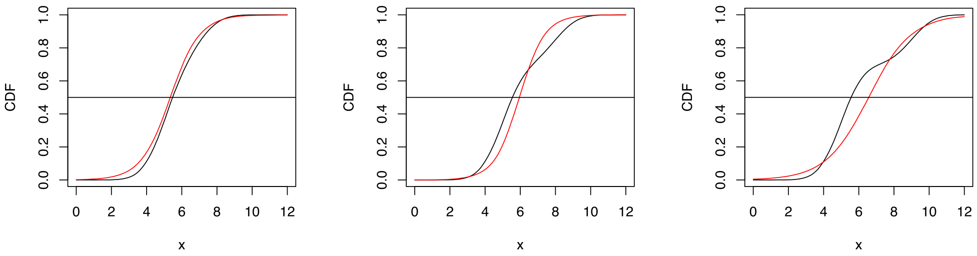
Degree 2. $\mu = 8$;

Degree 3. $\mu = 9$.

For the posited logistic regression model, we follow Robert and Casella (2004, Example 7.11) and take a default prior distribution for (β_0, β_1) to be

$$\pi(\beta_0, \beta_1) = \hat{b}^{-1} \exp\{\beta_0 - \hat{b}^{-1} e^{\beta_0}\},$$

which is simply a flat prior for β_1 and an exponential prior for e^{β_0} with scale $\hat{b} = \exp(\hat{\beta}_0 + \gamma)$, where $\hat{\beta}_0$ is the maximum likelihood estimator and $\gamma \approx 0.5772$ is Euler's



(a) Degree 1 (b) Degree 2 (c) Degree 3

Figure 4: Distribution function of the three normal mixtures (black) with the logistic distribution function (red) overlaid.

constant. Other kinds of prior distributions can be considered, but more sophisticated choices can lead to substantially longer computation times; see below.

The goal is, as in the previous section, to investigate the extent to which the learning rate selection methods can help the generalized Bayes posterior distribution to overcome the model misspecification, and Table 5 summarizes the results. There we present the average learning rate value, the coverage probability of 95% credible intervals for θ, the average length of those credible intervals, and mean square error, all based on 500 replications, for each pair of μ and sample size n. Here we see that, in the Degree 1 case where misspecification is relatively mild, the methods perform reasonably well in terms of coverage probability, but things get worse as sample size increases, a symptom of the model misspecification bias. For the Degree 2–3 cases with even more model misspecification, all the methods perform quite poorly. Apparently none of the learning rate selection methods can help the posterior overcome the relatively severe model misspecification bias in this example.

As mentioned above, there is an alternative default prior that is commonly used in logistic regression, namely, the Pólya–gamma prior of Polson et al. (2013). The challenge is that embedding a more sophisticated posterior sampling scheme, which involves the introduction of latent variables, inside the learning rate selection procedures is very expensive. Indeed, in our simulations, GPC could be used to tune the learning rate for a single data set, with $n = 100$, in about 4 minutes; SafeBayes, however, took about 5 times as long. Given these computational challenges, what we found in our limited simulation studies is that the Pólya–gamma prior does seem to have slightly better performance in the simulations. But, the overall message does not change, namely, that no learning rate selection method can help the generalized Bayes posterior adjust for moderate to extreme misspecification in the logistic regression example, at least not when the MCID is the inferential target.

Degree	n	Method	$\hat{\eta}$	Coverage	Length	MSE
1	100	GPC	0.904	0.938	0.883	0.051
		SafeBayes	0.790	0.953	0.987	0.053
		Holmes and Walker	0.999	0.927	0.834	0.051
		Lyddon et al.	1.003	0.923	0.830	0.051
	200	GPC	0.977	0.914	0.575	0.029
		SafeBayes	0.913	0.916	0.599	0.029
		Holmes and Walker	0.999	0.902	0.568	0.030
		Lyddon et al.	1.003	0.923	0.830	0.030
	400	GPC	0.910	0.912	0.418	0.015
		SafeBayes	0.822	0.926	0.450	0.015
		Holmes and Walker	0.999	0.890	0.397	0.015
		Lyddon et al.	0.988	0.892	0.401	0.015
2	100	GPC	0.786	0.866	1.148	0.139
		SafeBayes	0.890	0.893	1.283	0.138
		Holmes and Walker	1.000	0.836	1.071	0.137
		Lyddon et al.	0.986	0.838	1.085	0.137
	200	GPC	0.970	0.788	0.752	0.083
		SafeBayes	0.906	0.816	0.786	0.082
		Holmes and Walker	1.001	0.776	0.741	0.082
		Lyddon et al.	0.974	0.727	0.742	0.087
	400	GPC	0.901	0.622	0.539	0.067
		SafeBayes	0.833	0.646	0.574	0.067
		Holmes and Walker	0.999	0.564	0.511	0.067
		Lyddon et al.	0.969	0.588	0.518	0.067
3	100	GPC	0.955	0.742	1.412	0.345
		SafeBayes	0.891	0.750	1.482	0.346
		Holmes and Walker	1.001	0.726	1.377	0.344
		Lyddon et al.	0.970	0.731	1.385	0.343
	200	GPC	0.951	0.532	0.976	0.266
		SafeBayes	0.887	0.562	1.024	0.264
		Holmes and Walker	1.004	0.502	0.946	0.265
		Lyddon et al.	0.955	0.510	0.963	0.261
	400	GPC	0.820	0.244	0.693	0.247
		SafeBayes	0.893	0.290	0.743	0.248
		Holmes and Walker	1.000	0.228	0.657	0.248
		Lyddon et al.	0.951	0.236	0.671	0.247

Table 5: Summary of learning rate selection method performance in the misspecified logistic regression example based on 500 replications.

6.3 A Gibbs posterior

The generalized Bayes posterior is not able to overcome this apparently rather severe form of misspecification bias. As an alternative, we can consider a different type of posterior construction, the so-called *Gibbs posterior*. The generalized Bayes approach, which is model-based, proceeds as follows:

1. decide on the functional form that relates the diagnostic measure x to the probability that a patient reports feeling better ($y = +1$); this form could be quite flexible, e.g., with a nonparametric model;

2. choose priors for the parameters of that model;

3. fit that model, possibly with an adjusted learning rate;

4. find the marginal posterior for the MCID, θ, from the posterior of the model parameters.

The Gibbs posterior framework, on the other hand, constructs a posterior for θ *directly*, i.e., without introducing a functional form and without marginalization. This is done via a useful characterization of θ as the minimizer of a suitable expected loss; see below. More generally, the Gibbs posterior framework can be very effective when, like the present MCID application, the quantity of interest is not naturally understood as a model parameter that shows up in a likelihood function.

Define the loss function $\ell_\theta(x, y) = \frac{1}{2}\{1 - y\,\text{sign}(x - \theta)\}$ and the corresponding risk (expected loss) $R(\vartheta) = P\ell_\vartheta$. As Hedayat et al. (2015), showed, the MCID is the minimizer of R, i.e., $\theta^\star = \arg\min_\theta R(\theta)$. So the goal is to construct an empirical version of the risk function, and then a sort of posterior distribution that will concentrate around values that make the empirical risk small. For the empirical risk, let

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i, Y_i).$$

Then the Gibbs posterior distribution for θ has a density function defined as

$$\pi_n^{(\eta)}(\theta) \propto e^{-\eta n R_n(\theta)}\,\pi(\theta),$$

where $\eta > 0$ is, as before, the learning rate. In principle, all the different learning rate selection methods considered above can be applied to the Gibbs posterior framework to choose an appropriate value of η. Here, however, the loss function is not differentiable, which creates a challenge for the methods of Holmes and Walker (2017) and Lyddon et al. (2019). Therefore, in what follows, we only compare GPC and SafeBayes.

The results in Table 6 are to illustrate the performance of finding the learning rate using Gibbs posterior with scaling algorithm in Syring and Martin (2019) and Grünwald (2018). Since the true MCID is almost certain to be in the range of observed X values, the results here for both GPC and SafeBayes are based on uniform prior on $[X_{(1)}, X_{(n)}]$,

Degree	n	Method	$\hat{\eta}$	Coverage	Length	MSE
1	100	GPC	0.496	0.948	1.579	0.103
		SafeBayes	0.982	0.810	0.910	0.111
	200	GPC	0.396	0.932	1.177	0.060
		SafeBayes	0.986	0.700	0.587	0.073
	400	GPC	0.292	0.952	0.971	0.033
		SafeBayes	0.975	0.588	0.378	0.048
2	100	GPC	0.444	0.972	2.251	0.228
		SafeBayes	0.966	0.830	1.216	0.201
	200	GPC	0.339	0.950	1.797	0.143
		SafeBayes	0.967	0.700	0.750	0.123
	400	GPC	0.246	0.970	1.396	0.073
		SafeBayes	0.964	0.592	0.490	0.066
3	100	GPC	0.408	0.966	3.076	0.548
		SafeBayes	0.953	0.804	1.536	0.372
	200	GPC	0.313	0.958	2.452	0.351
		SafeBayes	0.964	0.692	0.892	0.205
	400	GPC	0.231	0.964	1.953	0.187
		SafeBayes	0.965	0.618	0.544	0.080

Table 6: Summary of GPC and SafeBayes learning rate selection method performance using a Gibbs posterior, based on 500 replications.

the sample range. Here we observe that the GPC is able to choose the learning rate such that the desired 95% coverage target for each sample size. SafeBayes, on the other hand, tends to choose too large of a learning rate, leading to (sometimes severe) under-coverage. In terms of computation time, with $n = 100$, GPC and SafeBayes took roughly 6 and 14 seconds per replication, respectively.

7 Conclusion

This paper investigated the performance of several existing procedures for choosing the learning rate parameter in generalized Bayes models. Our goal was to see which, if any, of these methods, are able to overcome the model misspecification bias and give valid posterior uncertainty quantification. While there are some models that are too severely misspecified for a learning rate adjustment alone to accommodate, we did find that such adjustments can be successful when misspecification is mild to moderate.

A take-away message is that, among the learning rate selection methods considered here, the GPC algorithm of Syring and Martin (2019) seems to be best suited overall for calibrating the generalized Bayes credible regions. This is not surprising, given that is precisely what the GPC algorithm is designed to do. GPC is computationally more expensive than, say, the method of Lyddon et al. (2019), but our results here suggest that the extra time/effort is well spent. Although GPC has been shown to have very good empirical performance here and in a number of other references, and the intuition behind why it *should* work is clear, there is still no formal proof that it does indeed provide valid posterior uncertainty quantification.

Finally, our focus here was exclusively on inference, but it would be of interest to see if/how different learning rate selection methods might assist in generalized Bayes prediction. After all, the prediction problem is one where it is possible to perform well even without a model, so developing a learning rate selection method that would correct for certain kinds of model misspecification, e.g., misspecified tails, should be within reach. That is, can a suitable choice of learning rate ensure that quantiles of the posterior predictive distribution achieve the nominal prediction coverage probability? In a recent manuscript (Wu and Martin, 2021), we show that it is possible to extend the GPC procedure to the prediction case, to achieve both valid and efficient prediction intervals across a wide range of applications, including those with spatial dependence.

References

Barron, A., Schervish, M. J., and Wasserman, L. (1999). "The consistency of posterior distributions in nonparametric problems." *Annals of Statistics*, 27(2): 536–561. MR1714718. doi: https://doi.org/10.1214/aos/1018031206. 109

Berk, R. H. (1966). "Limiting behavior of posterior distributions when the model is incorrect." *The Annals of Mathematical Statistics*, 37(1): 51–58. MR0189176. doi: https://doi.org/10.1214/aoms/1177699477. 105

Bhattacharya, A., Pati, D., and Yang, Y. (2019). "Bayesian fractional posteriors." *Annals of Statistics*, 47(1): 39–66. MR3909926. doi: https://doi.org/10.1214/18-AOS1712. 109

Bhattacharya, I. and Martin, R. (2022). "Gibbs posterior inference on multivariate quantiles." *Journal of Statistical Planning and Inference*, 218: 106–121. MR4337837. doi: https://doi.org/10.1016/j.jspi.2021.10.003. 109

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78(5): 1103–1130. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 106, 109, 110

Bunke, O. and Milhaud, X. (1998). "Asymptotic behavior of Bayes estimates under possibly incorrect models." *Annals of Statistics*, 26(2): 617–644. MR1626075. doi: https://doi.org/10.1214/aos/1028144851. 105, 108

De Blasi, P. and Walker, S. G. (2013). "Bayesian asymptotics with misspecified models." *Statistica Sinica*, 23: 169–187. MR3076163. 105

Diaconis, P. and Freedman, D. (1986a). "On inconsistent Bayes estimates of location." *Annals of Statistics*, 14(1): 68–87. MR0829556. doi: https://doi.org/10.1214/aos/1176349843. 105

Diaconis, P. and Freedman, D. (1986b). "On the consistency of Bayes estimates." *An-*

nals of Statistics, 14(1): 1–26. MR0829555. doi: https://doi.org/10.1214/aos/1176349830. 105

Faraway, J. (2016). *faraway: Functions and Datasets for Books by Julian Faraway*. R package version 1.0.7. URL https://CRAN.R-project.org/package=faraway 120

Fraser, D. A. S. (2011). "Is Bayes posterior just quick and dirty confidence?" *Statistical Science*, 26(3): 299–316. MR2918001. doi: https://doi.org/10.1214/11-STS352. 106

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *Annals of Statistics*, 27(1): 143–158. MR1701105. doi: https://doi.org/10.1214/aos/1018031105. 109

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28(2): 500–531. MR1790007. doi: https://doi.org/10.1214/aos/1016218228. 109

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. MR3587782. doi: https://doi.org/10.1017/9781139029834. 109

Grünwald, P. (2012). "The safe Bayesian: learning the learning rate via the mixability gap." In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, 169–183. Springer, Heidelberg. MR3042889. doi: https://doi.org/10.1007/978-3-642-34106-9_16. 106, 110

Grünwald, P. (2018). "Safe probability." *Journal of Statistical Planning and Inference*, 195: 47–63. MR3760837. doi: https://doi.org/10.1016/j.jspi.2017.09.014. 106, 126

Grünwald, P. and van Ommen, T. (2017). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it." *Bayesian Analysis*, 12(4): 1069–1103. MR3724979. doi: https://doi.org/10.1214/17-BA1085. 105, 106, 107, 109, 110, 116

Grünwald, P. D. and Mehta, N. A. (2020). "Fast rates for general unbounded loss functions: from ERM to generalized Bayes." *Journal of Machine Learning Research*, 21(56): 1–80. MR4095335. 109

Hedayat, A., Wang, J., and Xu, T. (2015). "Minimum clinically important difference in medical studies." *Biometrics*, 71(1): 33–41. MR3335347. doi: https://doi.org/10.1111/biom.12251. 122, 126

de Heide, R., Kirichenko, A., Grünwald, P., and Mehta, N. (2020). "Safe-Bayesian generalized linear regression." In *International Conference on Artificial Intelligence and Statistics*, 2623–2633. PMLR. 106, 113

Holmes, C. C. and Walker, S. G. (2017). "Assigning a value to a power likelihood in a general Bayesian model." *Biometrika*, 104(2): 497–503. MR3698270. doi: https://doi.org/10.1093/biomet/asx010. 106, 109, 111, 126

Huber, P. J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 221–233. University of California Press. MR0216620. 108

Kleijn, B. J. K. and van der Vaart, A. W. (2006). "Misspecification in infinite-dimensional Bayesian statistics." *Annals of Statistics*, 34(2): 837–877. MR2283395. doi: https://doi.org/10.1214/009053606000000029. 105

Kleijn, B. J. K. and van der Vaart, A. W. (2012). "The Bernstein-von-Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381. MR2988412. doi: https://doi.org/10.1214/12-EJS675. 105, 106, 108

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). "General Bayesian updating and the loss-likelihood bootstrap." *Biometrika*, 106(2): 465–478. MR3949315. doi: https://doi.org/10.1093/biomet/asz006. 106, 109, 111, 126, 127

Martin, R. (2017). "Invited comment on the article by van der Pas, Szabó, and van der Vaart." *Bayesian Analysis*, 12(4): 1254–1258. MR3983322. doi: https://doi.org/10.1214/19-STS707. 109

Martin, R. (2019). "False confidence, non-additive beliefs, and valid statistical inference." *International Journal of Approximate Reasoning*, 113: 39–73. MR3979518. doi: https://doi.org/10.1016/j.ijar.2019.06.005. 106

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 109

Martin, R. and Ning, B. (2020). "Empirical priors and coverage of posterior credible sets in a sparse normal mean model." *Sankhyā. Series A*, 82: 477–498. Special issue in memory of Jayanta K. Ghosh. MR4136243. doi: https://doi.org/10.1007/s13171-019-00189-w. 109

Martin, R. and Tang, Y. (2020). "Empirical priors for prediction in sparse high-dimensional linear regression." *Journal of Machine Learning Research*, 21(144): 1–30. MR4138128. 109

Martin, R. and Walker, S. G. (2019). "Data-dependent priors and their posterior concentration rates." *Electronic Journal of Statistics*, 13(2): 3049–3081. MR4010592. doi: https://doi.org/10.1214/19-ejs1600. 109

Miller, J. W. and Dunson, D. B. (2019). "Robust Bayesian inference via coarsening." *Journal of the American Statistical Association*, 114(527): 1113–1125. MR4011766. doi: https://doi.org/10.1080/01621459.2018.1469995. 106

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 56(1): 3–26. MR1257793. 111

Pérez-Rodríguez, P., Acosta-Pech, R., Pérez-Elizalde, S., Cruz, C. V., Espinosa, J. S., and Crossa, J. (2018). “A Bayesian genomic regression model with skew normal random errors.” *G3: Genes, Genomes, Genetics*, 8(5): 1771–1785. 120

Polson, N. G., Scott, J. G., and Windle, J. (2013). “Bayesian inference for logistic models using Pólya-Gamma latent variables.” *Journal of the American Statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 124

Ramamoorthi, R. V., Sriram, K., and Martin, R. (2015). “On posterior concentration in misspecified models.” *Bayesian Analysis*, 10: 759–789. MR3432239. doi: https://doi.org/10.1214/15-BA941. 105

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer, 2nd edition. MR2080278. doi: https://doi.org/10.1007/978-1-4757-4145-2. 123

Syring, N. and Martin, R. (2017). “Gibbs posterior inference on the minimum clinically important difference.” *Journal of Statistical Planning and Inference*, 187: 67–77. MR3638043. doi: https://doi.org/10.1016/j.jspi.2017.03.001. 109

Syring, N. and Martin, R. (2019). “Calibrating general posterior credible regions.” *Biometrika*, 106(2): 479–486. MR3949316. doi: https://doi.org/10.1093/biomet/asy054. 107, 109, 112, 113, 126, 127

Syring, N. and Martin, R. (2020a). “Gibbs posterior concentration rates under sub-exponential type losses.” arXiv:2012.04505. 109

Syring, N. and Martin, R. (2020b). “Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image.” *Annals of Statistics*, 48(3): 1498–1513. MR4124332. doi: https://doi.org/10.1214/19-AOS1856. 109

van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 108

Walker, S. and Hjort, N. (2001). “On Bayesian consistency.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 63(4): 811–821. MR1872068. doi: https://doi.org/10.1111/1467-9868.00314. 106, 108

Walker, S. G. (2013). “Bayesian inference with misspecified models.” *Journal of Statistical Planning and Inference*, 143(10): 1621–1633. MR3082220. doi: https://doi.org/10.1016/j.jspi.2013.05.013. 105

Walker, S. G., Lijoi, A., and Prünster, I. (2005). “Data tracking and the understanding of Bayesian consistency.” *Biometrika*, 92(4): 765–778. MR2234184. doi: https://doi.org/10.1093/biomet/92.4.765. 109

Wang, Z. and Martin, R. (2020). “Model-free posterior inference on the area under the receiver operating characteristic curve.” *Journal of Statistical Planning and Inference*, 209: 174–186. MR4096262. doi: https://doi.org/10.1016/j.jspi.2020.03.008. 109

Wu, P.-S. and Martin, R. (2021). “Calibrating generalized predictive distributions.” arXiv:2107.01688. 128

Zhang, T. (2006). "From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation." *Annals of Statistics*, 34(5): 2180–2210. MR2291497. doi: https://doi.org/10.1214/009053606000000704. 106, 109

Acknowledgments

The authors are grateful for the helpful suggestions from three anonymous reviewers on a previous version of the manuscript.