Calibrating general posterior credible regions

Nicholas Syring and Ryan Martin
Department of Statistics
North Carolina State University
(nasyring, rgmarti3)@ncsu.edu
April 27, 2017

Abstract

An advantage of methods that base inference on a posterior distribution is that uncertainty quantification, in the form of credible regions, is readily obtained. Except in well-specified situations, however, there is no guarantee that these credible regions will be calibrated in the sense that they achieve the nominal frequentist coverage probability, even approximately. To overcome this difficulty, we propose a general strategy—applicable to Bayes, Gibbs, and variational Bayes posteriors, among others—that introduces an additional scalar tuning parameter to control the posterior spread, and we develop an algorithm that chooses this parameter so that the corresponding credible region achieves the nominal coverage probability. Simulation results demonstrate that the proposed algorithm yields highly efficient credible regions in a variety of applications compared to existing methods.

Keywords and phrases: Bootstrap; coverage probability; Gibbs model; misspecified model; variational Bayes.

1 Introduction

An advantage of Bayesian and other more general Bayesian-like methods that base their inference on a suitable posterior distribution is that uncertainty quantification, in the form of credible regions for the unknown parameters, is readily available. For this uncertainty quantification to be meaningful, it is common to require that the specified credibility level agrees, at least approximately, with the frequentist coverage probability, i.e., that the 95% credibility regions read off from the posterior are approximately 95% confidence regions. In this case, we say that the posterior credible region is calibrated. For well-specified Bayesian models, one often has a Bernstein–von Mises theorem available to justify a calibration claim, but when the model is misspecified in at least one of several possible ways, calibration often fails. For example, Kleijn and van der Vaart (2012) derived a Bernstein–von Mises theorem for Bayesian posteriors under model misspecification, and pointed out that, even if concentration target and rate are correct, misspecification can still cause a lack of calibration; see page 362 in their paper and Section 2 below. Similarly, the commonly used variational Bayes posteriors (e.g., Jaakola and Jordan 1997; Jordan...
et al. 1999) often lack the desired calibration property, and correcting this is listed as one of the important open problems in Blei et al. (2017).

To address this problem, we propose to introduce, to the given posterior, an additional scalar tuning parameter, intended to control the spread of the posterior distribution. This formulation is inspired by the literature on Gibbs posteriors, where data and parameter of interest are connected via a loss function, instead of a likelihood; see, e.g., Bissiri et al. (2016), Alquier et al. (2016), Zhang (2006), Jiang and Tanner (2008), and Syring and Martin (2017). In such cases, a scale—or inverse temperature—parameter must be specified to properly weight the information in the data relative to that in the prior, but this ultimately boils down to tuning the Gibbs posterior spread. A similar formulation can be carried out for other Bayesian-like models, not just Gibbs posteriors; see Section 2.

Having introduced an extra parameter into the posterior, we then propose to select this tuning parameter such that the corresponding posterior credible regions are calibrated in the sense described above, and we present an algorithm, based on bootstrap and other Monte Carlo techniques, to implement this idea efficiently.

Similar questions about scaling posterior distributions to address one or more types of model misspecification have been considered recently in the literature. In particular, several ideas for choosing the posterior scaling are presented in Bissiri et al. (2016) and Holmes and Walker (2017), including hierarchical Bayes and loss/information matching, and Grünwald and Van Ommen (2016) propose to choose the scale parameter to minimize a type of prediction risk, similar in spirit to cross validation. These proposals are reasonable, but they do not provide any guarantees that the uncertainty quantification coming from the corresponding posterior distribution is meaningful. In contrast, our proposal here is designed specifically to make the corresponding posterior credible regions calibrated, at least approximately. The claimed calibration follows immediately from our construction, and the simulations presented in Section 4, covering several different models and types of posteriors, demonstrate the effectiveness of the proposed method.

The remainder of the paper is organized as follows. Section 2 sets our notation, defines our modified posterior distribution, with an extra calibration parameter, and explains the intuition behind our proposed approach. The general posterior calibration algorithm is presented in Section 3 and we discuss its basic properties. Section 4 contains several examples, including a Gibbs posterior in quantile regression, a misspecified Bayes posterior in linear regression, and a variational Bayes posterior in a mixture model, and Section 5 makes some concluding remarks.

2 Problem formulation

Suppose we have data $Z^n = (Z_1, \ldots, Z_n)$ consisting of iid observations from a distribution $P$; here, each $Z_i$ could be a vector or even a response–predictor variable pair, i.e., $Z_i = (X_i, Y_i)$. The quantity of interest is a parameter $\theta = \theta(P)$, a feature of the underlying distribution $P$, taking values in $\Theta$. Consider the following general construction of a posterior distribution for inference on $\theta$.

- Connect data $Z^n$ to a full set of parameters $\eta$ through either a statistical model for $P$, as in Bayes or other likelihood-based settings, or a suitable loss function, as in Gibbs or M-estimation settings.
• Introduce a prior $\Pi$ for the full parameter $\eta$, and a scale $\omega > 0$ to weight the information about $\eta$ in the data with that in the prior.

• Combine the prior, scale, and likelihood/loss to get a posterior distribution for $\eta$.

• Integrate to get the corresponding marginal posterior for $\theta$, denoted by $\Pi_{n,\omega}$.

This general recipe includes both the Bayesian and Gibbs posterior procedure, as well as variational Bayes, as we demonstrate in Section 4. It also covers classical empirical Bayes or other posteriors based on data-dependent priors (e.g., Fraser et al. 2010, Hannig et al. 2016, Martin and Walker 2016). The one technical requirement we have is that the posterior $\Pi_{n,\omega}$ be consistent in the sense that it concentrates, asymptotically, on the actual value $\theta(P)$ for each fixed $\omega$. Consistency must be verified case-by-case, but this is standard; see Section 4. Given that the posterior $\Pi_{n,\omega}$ is approximately centered around $\theta^*$, the use of credible regions to quantify uncertainty is reasonable.

For concreteness, consider the problem of estimating the median $\theta$ of a distribution $P$; more complicated examples are presented in Section 4. The median can be defined as the minimizer of the risk $R(\theta) = P_{\theta}\ell_\theta$, the expected value of the loss $\ell_\theta(z) = |z - \theta|$, under $P$. This loss forms a connection between $Z^n$ and $\theta$ and a Gibbs posterior is defined as

$$\Pi_{n,\omega}(d\theta) \propto e^{-\omega n R_n(\theta)} \Pi(d\theta),$$

where $R_n(\theta) = \mathbb{P}_n \ell_\theta$ is the empirical version of the risk, $\omega > 0$ is a scale parameter, and $\Pi$ is a prior for $\theta$. As an alternative, a Bayesian might specify a statistical model, such as $\text{Gamma}(\alpha, \beta)$, with likelihood $L_n(\eta)$ expressed in terms of parameters $\eta = (\alpha, \beta)$, and a prior $\Pi$ for $\eta$, and define a (scaled) posterior for $\theta$ as

$$\Pi_{n,\omega}(A) \propto \int_{\{\eta : F^{-1}_{\eta}(1/2) \in A\}} L_n(\eta)^\omega \Pi(d\eta),$$

where $F_{\eta}$ denotes the $\text{Gamma}(\alpha, \beta)$ distribution function. The choice between these two approaches, or variations thereof, depends largely on the willingness of the data analyst to specify a full model as well as on the goals of the analysis; the Gibbs approach provides inference on the median but nothing else, with minimal modeling assumptions, whereas the Bayes approach provides inference on virtually any feature, but with higher modeling and computational costs. In either case, the choice of scale $\omega$ is important.

Our proposed choice of scale is based on calibrating the posterior credible regions to be used for uncertainty quantification. Fix a level $\alpha \in (0, 1)$ and, for concreteness, consider the highest posterior density credible regions defined as

$$C_{\omega,\alpha}(Z^n) = \{\theta : \pi_{n,\omega}(\theta) \geq c_\alpha\},$$

where $\pi_{n,\omega}$ is the density function corresponding to the posterior $\Pi_{n,\omega}$, and $c_\alpha$ is a constant chosen so that the $\Pi_{n,\omega}$-probability assigned to $C_{\omega,\alpha}(Z^n)$ is equal to $1 - \alpha$. The scale parameter $\omega$ controls the spread of the posterior and, thereby, the size of these credible regions. Our proposal is to choose $\omega$ so that the credible regions are of the appropriate size to be calibrated, i.e., so that their coverage probability, $P\{C_{\omega,\alpha}(Z^n) \ni \theta(P)\}$, is approximately equal to $1 - \alpha$; see Section 3.

To better understand our proposal, recall that, in the classical setting of a well-specified Bayesian model with suitable regularity, the credible region will be calibrated,
Figure 1: Contours of the asymptotic distribution of the M-estimator (solid) and those of the asymptotic Gibbs posterior (dashed). Left: $\omega = 1$; Right: $\omega = 1/4$.

at least asymptotically, when $\omega = 1$. There has been recent interest in the misspecified case and, in particular, Kleijn and van der Vaart (2012) showed that even if a Bernstein–von Mises theorem holds, the posterior credible regions might not be calibrated. Roughly speaking, misspecification affects the shape of the posterior contours, which may be the wrong shape compared to the sampling distribution of the corresponding M-estimator. Varying the scale parameter can provide a conservative solution to this problem: we can stretch the posterior contours enough that they contain a differently-shaped but suitably calibrated confidence region; see Figure 1 for an illustration and Remark 2 below.

3 Posterior calibration algorithm

As discussed previously, our goal is to select the calibration parameter $\omega$ such that the corresponding posterior credible region are calibrated in the sense that the credibility level agrees, at least approximately, with the coverage probability. To this end, for our desired significance level $\alpha \in (0, 1)$, and our preferred credible region $C_{\omega, \alpha}(Z^n)$ as in (2), define the coverage probability function

$$c_{\alpha}(\omega \mid P) = P\{C_{\omega, \alpha}(Z^n) \ni \theta(P)\},$$

i.e., the $P$-probability that the credible region $C_{\omega, \alpha}(Z^n)$ contains the target $\theta(P)$. Then calibration requires that $\omega$ be such that

$$c_{\alpha}(\omega \mid P) = 1 - \alpha,$$

i.e., that the $100(1-\alpha)$% posterior credible region is also a $100(1-\alpha)$% confidence region. Of course, in practice, we cannot solve this equation because we do not know $P$. The approach described below is designed to get around this practical roadblock. Before we proceed, note that solving (3) is a fixed-$n$ exercise, so our aim is to get exact calibration in finite samples. Asymptotic approximations come into play, however, because $P$ is
unknown in real applications, but the numerical illustrations in Section 4 demonstrate
that we are, in fact, able to achieve exact calibration, at least in some cases.

To build up our intuition, start by assuming that \( P \) is known; later we will switch to
the more realistic case of unknown \( P \). Even in this unrealistic case, it is generally not
possible to solve for \( \omega \) in (3) explicitly, so numerical methods are required. It is possible to
solve (3) via stochastic approximation (Blei et al. 2017; Kushner and Yin 2003; Robbins
and Monro 1951) by iterating according to the rule

\[
\omega(t+1) = \omega(t) + \kappa_t \{ \hat{c}_\alpha(\omega(t) \mid P) - (1 - \alpha) \}, \quad t \geq 0
\]

(4)

where \( \hat{c}_\alpha(\omega \mid P) \) is a Monte Carlo approximation to the coverage probability, obtained
by simulating new copies of the data \( Z^n \) from \( P \), and \( (\kappa_t) \) is a non-stochastic sequence
such that \( \sum_t \kappa_t = \infty \) and \( \sum_t \kappa_t^2 < \infty \). For our numerical results in Section 4
we use \( \kappa_t = (t + 1)^{-3/4} \). In this case, if \( c_\alpha(\omega \mid P) \) is continuous and monotone decreasing in
\( \omega \), both very reasonable assumptions, then it follows from the almost supermartingale
convergence theorem of Robbins and Siegmund (1971), that \( \omega(t) \to \omega^* \) \( P \)-almost surely,
as \( t \to \infty \), where \( \omega^* \) is the solution to (3).

For the realistic case where \( P \) is unknown, the proposed approach changes in two
ways. First, since it is not possible to sample new copies of \( Z^n \) from \( P \), we replace
simulation from \( P \) with simulation from \( P_n \), i.e., we sample with replacement from the
observed data \( Z^n \). Second, since we also do not know \( \theta(P) \), we cannot check if a given
credible region \( C_{\omega,\alpha}(Z^n) \) covers it. Instead, we use \( \theta(P_n) \) in place of \( \theta(P) \). This results
in an empirical version of the coverage probability \( c_\alpha(\omega \mid P) \), namely,

\[
c_\alpha(\omega \mid P_n) = P_n \{ C_{\omega,\alpha}(Z^n) \ni \theta(P_n) \},
\]

(5)

and our proposal is to find \( \omega \) such that

\[
c_\alpha(\omega \mid P_n) = 1 - \alpha.
\]

(6)

In practice, we cannot evaluate \( c_\alpha(\omega \mid P_n) \) either, but bootstrap will provide a Monte
Carlo estimator, which we denote by \( \hat{c}_\alpha(\omega \mid P_n) \). We can proceed to solve (3) by using
the same stochastic approximation procedure described above for the known-\( P \) case.
Collectively, these steps to solve this equation make up our general posterior calibration
(GPC) algorithm. An R code implementation for each of the examples in Section 4 is
available at https://github.com/nasyring/GPC.

Remark 1. In most applications, the credible regions \( C_{\omega,\alpha}(Z^n) \) will not be available in
closed form, so posterior sampling will be needed in such cases. But despite having
several moving parts—bootstrap, MCMC, and stochastic approximation—the proposed
GPC algorithm is relatively inexpensive computationally. For example, in the quantile
regression problem in Section 4.1 with a two-dimensional parameter, sample size \( n = 100 \),
\( B = 200 \) bootstrap samples, and \( M = 2000 \) posterior samples, the algorithm took less
than 10 seconds to converge on a Windows desktop computer with a 4.0 GHz Intel Core
i7 processor. We believe that a minimal extra computational investment is a fair trade
for calibrated posterior credible regions.

Remark 2. The workhorses of the GPC algorithm, namely, bootstrap, stochastic approx-
imation, and MCMC, are widely used and, on their own, theoretically sound. But having
Algorithm 1 – General Posterior Calibration.

Fix a convergence tolerance $\varepsilon > 0$ and an initial guess $\omega^{(0)}$ of the calibration parameter. Take $B$ bootstrap samples $\tilde{Z}_1^n, \ldots, \tilde{Z}_B^n$ of size $n$. Set $t = 0$ and do:

1. Construct credible regions $C_{\omega^{(t)}, \alpha}(\tilde{Z}_b^n)$ for each $b = 1, \ldots, B$.
2. Evaluate the empirical coverage $\hat{c}_n(\omega^{(t)} | \mathbb{P}_n)$ as in (5).
3. If $|\hat{c}_n(\omega^{(t)} | \mathbb{P}_n) - (1 - \alpha)| < \varepsilon$, then stop and return $\omega^{(t)}$ as the output; otherwise, update $\omega^{(t)}$ to $\omega^{(t+1)}$ according to (4), set $t \leftarrow t + 1$, and go back to Step 1.

them all working together in tandem makes a general theoretical analysis of the GPC algorithm very challenging. Here, we provide only some technical insight as to why the algorithm works, leaving a complete theoretical analysis for future research.

Let the interest parameter be defined via a risk function $R(\theta) = P(\ell(\theta))$, so that $\theta(P) = \arg\min R(\theta)$. In this case, a Gibbs posterior $\Pi_{n, \omega}$ is defined as in (4), where $R_n(\theta) = \mathbb{P}_n \ell_\theta$ is the empirical risk. Under suitable regularity conditions, the Gibbs posterior will resemble a normal distribution, centered at $\theta(P_n) = \arg\min R_n(\theta)$, with asymptotic covariance matrix $\omega^{-1} \Sigma_n$, where $\Sigma_n = (n V_\theta(\mathbb{P}_n))^{-1}$ and $V_\theta$ is the second derivative matrix of $R(\theta)$. So, the credible region $C_{\omega, \alpha}(Z^n)$ will look roughly like

$$\{ \theta : \omega((\theta - \theta(\mathbb{P}_n))^\top \Sigma_n^{-1}(\theta - \theta(\mathbb{P}_n)) \leq \xi_\alpha \},$$

where $\xi_\alpha$ is the appropriate chi-square quantile. On the other hand, the asymptotic covariance matrix of $\theta(\mathbb{P}_n)$ is given by $\Psi_n = n^{-1} V^{-1}_\theta(\mathbb{P}) M V^{-1}_\theta(\mathbb{P})$, where $M = P \hat{\ell}_\theta(\mathbb{P}) \ell_\theta(\mathbb{P})$ and $\hat{\ell}_\theta$ is the derivative of $\theta \rightarrow \ell_\theta$, so an asymptotic confidence region is

$$\{ \theta : (\theta - \theta(\mathbb{P}_n))^\top \Psi_n^{-1}(\theta - \theta(\mathbb{P}_n)) \leq \xi_\alpha \}.$$

In general, $\omega \Sigma_n^{-1}$ and $\Psi_n^{-1}$ are different, so the credible region has a different shape than the confidence region and, therefore, may not be calibrated. But the GPC algorithm will take $\omega$ roughly equal to the smallest eigenvalue of $\Psi_n^{-1/2} \Sigma_n \Psi_n^{-1/2}$, so that the credible region contains the aforementioned confidence region, making the latter conservatively calibrated; exact calibration is possible only if $\Psi_n = c \Sigma_n$ for some scalar $c > 0$. Since $\Psi_n^{-1/2} \Sigma_n \Psi_n^{-1/2}$ has a limit as $n \rightarrow \infty$, we expect that our GPC solution to (5) will converge to the smallest eigenvalue of that limiting matrix, hence asymptotic calibration.

For a quick proof-of-concept, suppose the data $Z^n$ are iid $P$ and the population mean $\theta = \int z \, dP(z)$ is the quantity of interest. Here we take $P = \mathcal{N}(0, 1)$, so that $\theta = 0$. We consider three posterior distributions: a Bayes model using the correct normal likelihood; a Gibbs posterior using $R_n(\theta) = \sum_{i=1}^n (Z_i - \theta)^2$, and a misspecified Bayesian posterior with a Laplace likelihood. For the well-specified Bayes and the Gibbs posteriors, we expect the GPC algorithm to select $\omega \approx 1$ and $\omega \approx 0.5$, respectively; for the Laplace model, based on the $V_\theta$ and $M$ calculations in Remark 2 we expect $\omega \approx 0.64$. Figure 3 plots the mean trajectories of the $\omega$ values obtained from our algorithm, with error bars, as a function of $n$. These results confirm our expectations based on theory.
Figure 2: Mean choice of $\omega$ over 100 simulated standard normal data sets of sizes 100, 250, 500, and 1000 using the true likelihood, a Gibbs model, and a Laplace likelihood. Vertical bars represent two standard deviations from the mean.

4 Applications

4.1 Quantile regression

In quantile regression, for fixed $\tau \in (0, 1)$, we are interested in the $\tau$th quantile of the response $Y \in \mathbb{R}$, given the covariates $X \in \mathbb{R}^{p+1}$, expressed as

$$Q_\tau(Y \mid X) = X^\top \theta,$$

where dimension $p+1$ represents an intercept and $p$ covariates. In this formula, the vector $\theta$ depends on $\tau$ but, for notational simplicity, we will omit this dependence. This model specifies no parametric form for the conditional distribution of $Y$ given $X$. Inference on the quantile regression coefficient $\theta$ may be carried out using asymptotic approximations [Koenker 2005, Theorem 4.1] or by using the bootstrap [Horowitz 1998]. A Bayesian approach would also be attractive, but no distributional form for the conditional distribution is given in (7), hence no likelihood. A workaround that has been considered by several authors (e.g., Sriram 2015; Sriram et al. 2013; Yu and Moyeed 2001) is to use a (misspecified) asymmetric Laplace likelihood. This corresponds to a Gibbs model (1) using the empirical risk

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^n |(Y_i - X_i^\top \theta)(\tau - I_{Y_i - X_i^\top \theta < 0})|$$

based on the usual check-loss function, where $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$, are the observations, and $I$ is the indicator function.

It follows from [Kleijn and van der Vaart 2012] that the Gibbs posterior based on (8) satisfies a Bernstein–von Mises theorem. Despite the desirable convergence result, the variance mismatch discussed in Section 2 causes the credible regions to be too large and over-cover, a sign of inefficiency. On the other hand, the GPC algorithm calibrates the intervals exactly, for all $n$, without loss of efficiency in terms of interval lengths.
Table 1: Comparison of 95% posterior credible intervals of the median regression parameters from five methods: BEL.s; BDL; Normal; the confidence interval computed using the asymptotic normality of the M-estimator; $\omega \equiv 0.8$, the scaled posterior with $\omega$ fixed equal to 0.8; and GPC. Coverage probability and average interval lengths are computed over 5000 simulated data sets for our method, normal intervals, and fixed-\omega intervals. Results for BEL.s and BDL are taken from Yang and He (2012) and were calculated from 1000 simulated data sets.

To demonstrate this, we revisit a simulation example presented in Yang and He (2012). For $\tau = 0.5$, the model they consider is

$$Y_i = \theta_0 + \theta_1 X_i + e_i, \quad i = 1, \ldots, n,$$

where $\theta_0 = 2$, $\theta_1 = 1$, $e_i \overset{iid}{\sim} \mathcal{N}(0, 4)$, and $X_i \overset{iid}{\sim} \text{ChiSq}(2) - 2$. For this model, the authors showed numerically that their proposed Bayesian empirical likelihood approach (“BEL.s”) produced credible intervals with approximate coverage near the nominal 95% level. Moreover, compared to the Bayesian method with misspecified asymmetric Laplace likelihood (“BDL”) or, equivalently, our posterior with $\omega$ chosen by averaging residuals, their method is shown to be more efficient in terms of interval length. The results for these methods are presented in Table 1, along with the results from the posterior intervals scaled by the algorithm.

There are two key observations to be made. First, our method calibrates the credible intervals to have exact 95% coverage across the range of $n$, while the other methods tend to over-cover. Second, our credible intervals tend to be shorter than those of the other methods, especially for $n = 100$. All three methods have a $n^{-1/2}$ convergence rate so, for large $n$, we cannot expect to see substantial differences between the various methods. Therefore, the small-$n$ case should be the most important and, at least in this case, the credible intervals calibrated using our algorithm are clearly the best.

Finally, considering that in smooth models we expect $\omega$ to account for the difference in asymptotic variance between the posterior and the M-estimator, it is reasonable to ask if we need a calibration algorithm at all, i.e., can we get by with a fixed value of $\omega$ based on these asymptotic variances? A comparison of the asymptotic variance of the posterior with that of the M-estimator shows that $0.80\Sigma_n^{-1} \approx \Psi_n^{-1}$; therefore, we can take $\omega \equiv 0.80$ in an attempt to calibrate posterior credible intervals with a fixed scaling. Table 1 shows that our algorithm is still better than using a fixed scale based on asymptotic normality, especially at smaller sample sizes where the normal approximation is less justifiable.
4.2 Linear regression

Consider the usual multiple linear regression model for data \((X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}\)

\[
Y_i = \beta_0 + X_i^\top \beta + \sigma e_i, \quad i = 1, \ldots, n,
\]  

(9)

where \(\beta \in \mathbb{R}^p\) is the vector of slope coefficients, \(\sigma > 0\) is an unknown scale parameter, and \(e_1, \ldots, e_n\) are assumed to be iid \(\mathcal{N}(0, 1)\). Suppose, however, that the constant error variance assumption is violated, in particular, \(e_i \sim \mathcal{N}(0, \sigma^2 \|X_i\|)\), \(i = 1, \ldots, n\), independent.

Our choice of predictor-dependent variance is a less-stylized version of that in Grünwald and Van Ommen (2016). The proposed model is, therefore, misspecified, but our goal is still to obtain calibrated inference on \(\theta = (\beta_0, \beta)\).

The Jeffreys prior is a reasonable default choice with density \(\pi(\eta) \propto (\sigma^2)^{-3/2}\) (Ibrahim and Laud 1991) for the full parameter \(\eta = (\theta, \sigma^2)\). Since this prior is probability-matching for the location-scale model (e.g., Datta and Mukerjee 2004), we may expect that the posterior credible intervals would be approximately calibrated for our linear regression. However, for a misspecified model, calibration might fail; in fact, as shown in Table 2, the credible intervals are too narrow and tend to undercover.

To investigate the performance of our proposed posterior calibration method compared to several others, we carry out a simulation study. We simulated data sets of \(n = 50\) observations. Each \(X_i \in \mathbb{R}^3\) is multivariate normal with zero mean and unit variance for each element, and pairwise correlation 0.5 for \(X_{i1}\) and \(X_{i2}\) and zero otherwise. To sample \(Y_i\) we use \(\beta_0 = 0, \beta = (1, 2, -1)^\top\), and \(\sigma = 1\). Although the error variance contains \(\|X_i\|\), the regular tests for constant variance do not detect the heteroscedasticity. Table 2 shows the estimated coverage probability and mean lengths of several posterior credible intervals for the components of \(\theta\). Besides those scaled by the GPC algorithm, we consider a misspecified Bayes approach that fixes \(\omega \equiv 1\), and posteriors with scale \(\omega\) chosen by the method in Holmes and Walker (2017) and the SafeBayes method in Grünwald and Van Ommen (2016, Algorithm 1). The results in Table 2 show that for this example SafeBayes performs similarly to GPC, while the method in Holmes and Walker (2017) does not improve upon the misspecified Bayesian model in terms of calibration.

Figure 3 shows a boxplot comparison of the scale parameters chosen by the three posterior scaling methods for the misspecified Bayesian posterior. Our algorithm, along with the SafeBayes method, tends to produce smaller values of \(\omega\) than the method of Holmes and Walker. Small values of \(\omega\) mean higher posterior variance and wider credible intervals, which explains these method’s improvement in calibration. While both our algorithm and SafeBayes pick \(\omega \approx 0.8\) on average, the distribution of \(\omega\) is much more concentrated using our algorithm.

4.3 Variational inference for a normal mixture model

Variational inference offers a competing method to Markov chain Monte Carlo for approximating the posterior distribution. This approach specifies a family of distributions—often a normal family—as candidate posteriors and then chooses the parameters of that family to minimize the Kullback–Leibler divergence from the true posterior. The variational posterior is simple by construction and, if carefully chosen, will be consistent (e.g., Wang and Titterington 2005), but as noted in Blei et al. (2017), misspecification causes the variational posterior variance to be too small.
Table 2: Empirical coverage probabilities of 95% credible intervals and average interval lengths (and standard deviations) calculated using 5000 simulations from the model described in Section 4.2.

| Method                  | \( \beta_0 \)  | \( \beta_1 \)  | \( \beta_2 \)  | \( \beta_3 \)  |
|-------------------------|-----------------|-----------------|-----------------|-----------------|
| Misspecified Bayes      | 0.94            | 0.89            | 0.88            | 0.87            |
| length                  | 0.99(0.15)      | 1.16(0.20)      | 1.16(0.20)      | 1.01(0.17)      |
| GPC                     | 0.98            | 0.94            | 0.94            | 0.93            |
| length                  | 1.17(0.18)      | 1.36(0.23)      | 1.36(0.24)      | 1.18(0.20)      |
| SafeBayes               | 0.96            | 0.93            | 0.94            | 0.92            |
| length                  | 1.19(0.26)      | 1.40(0.31)      | 1.39(0.33)      | 1.21(0.28)      |
| Holmes and Walker       | 0.91            | 0.84            | 0.80            | 0.82            |
| length                  | 0.87(0.18)      | 1.01(0.22)      | 1.01(0.22)      | 0.87(0.18)      |

Figure 3: Boxplots of \( \omega \) for the model described in Section 4.2 using GPC, SafeBayes (Grunwald and Van Ommen 2016), and the method in Holmes and Walker (2017) over 5000 simulated data sets.

As an example, we consider the normal mixture model presented in Blei et al. (2017), i.e., \( Y_1, \ldots, Y_n \) are iid observations from the mixture model

\[
\sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2) .
\]

The full parameter \( \eta \) consists of the mixture weights \( \pi_1, \ldots, \pi_K \), means \( \mu_1, \ldots, \mu_K \), and variances \( \sigma_1^2, \ldots, \sigma_K^2 \), but we will consider inference only on the means. We can construct a variational posterior for \( \eta \) following Algorithm 2 in Blei et al. (2017), which approximates the posterior by a multivariate normal. The additional scale factor \( \omega \) in our modified variational posterior \( \Pi_{\eta, \omega} \) only adjusts the overall scale of this multivariate normal. Therefore, if \( m_1, \ldots, m_K \) and \( v_1, \ldots, v_K \) are the means and variances, respectively, of this variational posterior for the mixture means \( \mu_1, \ldots, \mu_K \), then the corresponding \( \omega \)-scaled variational posterior 100(1 – \( \alpha \))% credible intervals are of the form

\[
\mu_k \pm z_{\alpha/2} \omega v_k^{1/2} , \quad k = 1, \ldots, K .
\]
Table 3: Empirical coverage probability and average length (standard deviation) of the credible intervals for $(\mu_1, \mu_2)$ based on our GPC algorithm and the variational posterior (VI) in [Blei et al.] (2017) over 5000 simulated data sets from the mixture model (10).

|       | $\mu_1$ | $\mu_2$ |
|-------|---------|---------|
| **GPC** | coverage | 0.96 | 0.96 |
|       | length  | 0.67 (0.08) | 0.67 (0.08) |
| **VI**  | coverage | 0.92 | 0.92 |
|       | length  | 0.55 (0.03) | 0.55 (0.03) |

It is straightforward to incorporate this variational posterior setup into our GPC algorithm; the computational investment is in carrying out the optimization needed for the variational approximation at each bootstrap step, but then the credible intervals are available in closed-form so no posterior sampling is needed.

We claim that the GPC algorithm will properly scale the variational posterior, calibrating the corresponding credible intervals, correcting the under-estimation of variance noted in [Blei et al.] (2017). To demonstrate this, we carry out a simple simulation study. We take $K = 2$, $\pi_1 = \pi_2 = 1/2$, $(\mu_1, \mu_2) = (-2, 2)$, and $\sigma_1 = \sigma_2 = 1$. Table 3 shows the empirical coverage probabilities and mean lengths of the 95% credible intervals based on Algorithm 2 in [Blei et al.] (2017) and our GPC algorithm. Apparently, our GPC algorithm corrects the underestimated variance of the variational posterior, producing credible intervals that are slightly conservative.

5 Discussion

The sensitivity of Bayesian credible sets to the posited probability model makes obtaining calibrated inference a challenging problem. Our linear regression example demonstrates this sensitivity when we take the model for granted. However, misspecification can happen in a variety of settings, and not always unintentionally. In quantile regression, the model is determined by a risk function rather than a likelihood, making traditional Bayesian inference using the true likelihood elusive. And, other times, computational considerations make variational posteriors an attractive alternative to a fully Bayesian analysis. Our posterior calibration algorithm may provide a solution in all of these settings by correcting model misspecification to produce, at least approximately, calibrated inferences.

Although the focus in this paper is on misspecified models, it may still be desirable to apply our algorithm even when the true likelihood is used. The reason is that the our algorithm can aid in producing calibrated inferences for the given sample size, regardless of the prior distribution used. This facilitates the use of informative priors, if available, instead of default priors, while still gaining the desired calibration property.

Finally, while it is clear that the GPC algorithm produces approximately calibrated credible sets, a detailed theoretical study is needed. The techniques we have used—stochastic approximation, MCMC, and bootstrap—each are theoretically sound on their own, but very complicated when used in tandem. Further work in this direction may help provide guidance, but the lack of completely rigorous theory does not take away from the encouraging examples shown throughout the paper.
Acknowledgments

This work is partially supported by the U. S. Army Research Offices, Award #W911NF-15-1-0154.

References

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.*, 17:1–41.

Bissiri, P., Holmes, C., and Walker, S. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 78(5):1103–1130.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. To appear in: *J. Amer. Statist. Assoc*.

Datta, G. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Springer, New York.

Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(5):631–654.

Grunwald, P. and Van Ommen, T. (2016). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Unpublished manuscript, arXiv:1412.3730.

Hannig, J., Iyer, H., Lai, R., and Lee, T. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.*, 111:1346–1361.

Holmes, C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a Bayesian model. To appear in: *Biometrika*.

Horowitz, J. (1998). Bootstrap methods for median regression models. *Econometrica*, 66:1327–1351.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys’s prior. *J. Amer. Statist. Assoc.*, 86(416):981–986.

Jaakola, T. S. and Jordan, M. I. (1997). A variational approach to Bayesian logistic regression models and their extensions. *Sixth International Workshop on Artificial Intelligence and Statistics*, 82.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Stat.*, 36(5):2207–2231.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37:183–233.

Kleijn, B. and van der Vaart, A. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381.
Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs, volume 38. Cambridge Univ. Press, Cambridge.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, New York, second edition.

Martin, R. and Walker, S. (2016). Optimal Bayesian posterior concentration rates with empirical priors. Unpublished manuscript, arXiv:1604.05734.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. In Rustagi, J. S., editor, *Optimizing Methods in Statistics*, pages 233–258. Academic Press, New York.

Sriram, K. (2015). A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Stat. Probab. Lett.*, 107:18–26.

Sriram, K., Ramamoorthi, R. V., and Ghosh, P. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Anal.*, 8(2):479–504.

Syring, N. and Martin, R. (2017). Gibbs posterior inference on the minimum clinically important difference. *J. Stat. Plan. Inference*, 187:67–77.

Wang, B. and Titterington, D. M. (2005). Inadequacy of interval estimates corresponding to variational bayesian approximations. In *AISTATS*.

Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Ann. Stat.*, 40(2):1102–1131.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Stat. Probab. Lett.*, 54(4):437–447.

Zhang, T. (2006). From $\epsilon$-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Ann. Stat.*, 34(5):2180–2210.