

Gibbs posterior inference on multivariate quantiles

Indrabati Bhattacharya* and Ryan Martin†

Abstract

Bayesian and other likelihood-based methods require specification of a statistical model and may not be fully satisfactory for inference on quantities, such as quantiles, that are not naturally defined as model parameters. In this paper, we construct a direct and model-free Gibbs posterior distribution for multivariate quantiles. Being model-free means that inferences drawn from the Gibbs posterior are not subject to model misspecification bias, and being direct means that no priors for or marginalization over nuisance parameters are required. We show here that the Gibbs posterior enjoys a root- n convergence rate and a Bernstein–von Mises property, i.e., for large n , the Gibbs posterior distribution can be approximated by a Gaussian. Moreover, we present numerical results showing the validity and efficiency of credible sets derived from a suitably scaled Gibbs posterior.

Keywords and phrases: Bernstein–von Mises phenomenon; concentration rate; credible sets; learning rate; multivariate median.

1 Introduction

In multivariate analysis, often the quantity of interest is the multivariate mean vector. However, there are situations when the mean is not a very good measure of location, for example, when the data is skewed, has outliers, etc. In such situations, a multivariate median would be a much more robust measure of the distribution’s center. Unfortunately, there is no universally accepted definition of a multivariate median, because there is no objective basis of ordering the data points in higher dimensions. Over the years, various definitions of multivariate medians and, more generally, multivariate quantiles have been proposed; see Small (1990) for a comprehensive review.

The most common version of a multivariate median is called the ℓ_1 -median, which is characterized through an ℓ_1 -optimization problem. Define

$$\ell_\theta(x) = \|x - \theta\|_r - \|x\|_r,$$

where $\|x\|_r = (\sum_{j=1}^d |x_j|^r)^{1/r}$ is the usual ℓ_r -norm of a d -dimensional vector x in \mathbb{R}^d , for $r \in (1, \infty)$ a fixed constant. Following Small (1990), the ℓ_1 -median of the random vector $X \sim P$, taking values in \mathbb{R}^d , is

$$\theta(P) = \arg \min_{\theta} P \ell_\theta,$$

*Department of Biostatistics and Computational Biology, University of Rochester Medical Center, indrabati.bhattacharya@urmc.rochester.edu

†Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

where we use the notation $Pf = \int f dP$ to denote the expected value of a random variable $f(X)$ with respect to the distribution P . The special case $r = 2$ is called the spatial median and was studied in, e.g., Brown (1983). Additional details about ℓ_1 -medians and, more generally, about multivariate quantiles are given in Section 2.1 below.

In statistical applications, the distribution P is unknown, but an independent and identically distributed (iid) sample X_1, \dots, X_n of random vectors in \mathbb{R}^d are available from P . Based on the above formulation, an immediate strategy for estimating the ℓ_1 -median is to follow the definition of $\theta(P)$ but replace the distribution P with the empirical distribution, i.e., with $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x denotes a point-mass distribution at the point $x \in \mathbb{R}^k$. That is, the standard point estimate of $\theta(P)$ is

$$\hat{\theta}_n = \theta(\mathbb{P}_n) = \arg \min_{\theta} \mathbb{P}_n \ell_{\theta}. \quad (1)$$

The spatial sample median is a highly robust estimator of location, in particular, its breakdown point is 1/2. Also, Möttönen et al. (2010) investigated the asymptotic properties of spatial median and proved its asymptotic normality.

Beyond estimation, if the goal is probabilistic inference on a multivariate median or quantile, i.e., via a “posterior distribution,” then the chief difficulty is that these are not naturally described as parameters in a statistical model. That is, no standard or otherwise “reasonable” model for multivariate data will include a quantile in its parametrization, so some potentially dangerous non-linear marginalization (e.g., Fraser 2011; Martin 2019) would typically be required. More importantly, with specification of a model comes the risk of model misspecification bias, and, since our quantity of interest is well-defined without a model, it is not clear what can be gained by working in a model-based framework to balance out the risk of misspecification bias. However, there are advantages to having a posterior probability distribution on which to base inferences; for example, a posterior density provides a nice visual summary of uncertainty, credible regions can be immediately read off from it without asymptotic approximations, and prior information about the quantity of interest can be incorporated whenever it is available. A Bayesian approach that both achieves these desirable features (more or less) and avoids the risk of model misspecification bias must be nonparametric. That is, assign a prior distribution—say, a Dirichlet process (Ferguson 1973)—to the infinite-dimensional P , get the corresponding posterior, and then do the non-trivial marginalization to $\theta = \theta(P)$. We call this an *indirect* approach. Aside from computational challenges, a downside of the indirect approach is that incorporating available prior information about the quantile is difficult; probably the best option is to choose a Dirichlet process base measure to have quantile equal to a prior guess, but it is not clear how this (and other features of the specified base measure) affect the marginal posterior for the quantile.

Is it possible to develop a posterior for the multivariate quantile in a more *direct* way, without marginalization, etc.? Here we investigate the construction of a *Gibbs posterior* for a multivariate quantile. On one hand, like M-estimation, this approach uses a suitable loss function, rather than a likelihood, to connect the quantity of interest to the observed data, which eliminates the risk of model misspecification bias. On the other hand, like Bayesian inference, it produces a genuine posterior distribution and allows for the direct incorporation of prior information. After some background about multivariate quantiles and Gibbs posteriors in Section 2, we define our object of interest, namely, the Gibbs posterior distribution for a multivariate quantile, and investigate its properties. In

particular, in Section 3.2, we first establish that the Gibbs posterior concentrates around the true quantile at the usual root- n rate and, second, that it has an asymptotic Gaussian approximation in the Bernstein–von Mises sense. Unfortunately, the covariance matrix in this Gaussian approximation is “wrong” in the sense that it does not match that of the M-estimator around which it is centered. Fortunately, the Gibbs posterior depends on a user-specified *learning rate* (e.g., Bissiri et al. 2016; Grünwald and van Ommen 2017; Syring and Martin 2019) which can be tuned to at least partially correct for the covariance matrix mismatch. We use a bootstrap-based calibration algorithm proposed by Syring and Martin (2019) for choosing the learning rate, which we describe in Section 3.3. In Section 4.1, we compare the finite-sample performance of our proposed Gibbs posterior inference to that based on existing Bayesian approaches. The two key take-aways are: (a) the Gibbs posterior outperforms the model-based parametric Bayesian approach in misspecified situations, and (b) the Gibbs posterior credible sets are at least as good as the nonparametric Bayes credible sets in terms of coverage but they are more efficient in some cases. We also apply the Gibbs posterior approach to infer the spatial median based on a real data set in Section 4.2, for which the assumption of normality is debatable. We show that the Gibbs posterior outperforms a normality-based Bayesian solution in terms of out-of-sample risk, implying that our Gibbs solution avoids the inherent bias coming from the normality assumption. Some concluding remarks are given in Section 5 and proofs of the two main theorems are presented in the Appendix.

2 Background

2.1 Multivariate quantiles

Again, the lack of a well-defined ordering of multivariate observations creates a major issue in defining multivariate quantiles. Abdous and Theodorescu (1992) and Babu and Rao (1989) investigated the coordinate-wise medians and quantiles. However, the coordinate-wise quantiles do not provide much information about the joint distribution of the vector and they also lack some desirable geometric properties, namely, rotational invariance. To fill this gap, Chaudhuri (1996) introduced the notion of geometric quantiles based on the geometric configuration of multivariate data clouds. These quantiles are natural generalizations of the ℓ_1 -median. For univariate observations $X_1, \dots, X_n \in \mathbb{R}$, the sample α th quantile $\alpha \in (0, 1)$ is obtained by minimizing $\xi \mapsto \sum_{i=1}^n \{|X_i - \xi| + u(X_i - \xi)\}$, with $u = 2\alpha - 1$. Generalizing this idea to higher dimensions, the d -dimensional geometric quantiles, with ℓ_r -norm, are indexed by points in the open unit ball $B_q^{(d)} = \{u \in \mathbb{R}^d : \|u\|_q < 1\}$, where q is the Hölder conjugate of r , i.e., $r^{-1} + q^{-1} = 1$. Thus, for $u \in B_q^{(d)}$, the d -dimensional sample u -quantile is then defined as

$$\widehat{Q}_n(u) = \arg \min_{\xi \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \Phi_r(u, X_i - \xi), \quad (2)$$

where $\Phi_r(u, t) = \|t\|_r + \langle u, t \rangle$, with $\langle \cdot, \cdot \rangle$ being the usual inner product. It is easy to see that $\widehat{Q}_n(0)$ is the same as the ℓ_1 -median; $\widehat{\theta}_n$. The population analog of $\widehat{Q}_n(u)$ is given by

$$Q_P(u) = \arg \min_{\xi \in \mathbb{R}^d} P\{\Phi_r(u, X - \xi) - \Phi_r(u, X)\}. \quad (3)$$

Chaudhuri (1996) showed that the geometric quantiles are both equivariant under location transformation and homogeneous scale transformation of the individual coordinates.

Chaudhuri's approach has received considerable attention in the literature, and has also been extended to regression contexts, for example, in Chakraborty (1999, 2003). One other notable approach to generalizing univariate quantiles to multivariate case is the directional quantile approach developed by Hallin et al. (2010). A directional quantile τ is a function of two components, namely, a direction vector u and a depth $\gamma \in (0, 1)$. Then the $\tau = u\gamma$ directional quantile, denoted by λ_τ is a hyperplane through \mathbb{R}^d .

In a Bayesian setting, Bhattacharya and Ghosal (2020) considered the use of a Dirichlet process prior on the underlying distribution P , and explored properties of the corresponding marginal posterior distribution of $Q_P(u)$ or, more precisely, a Bayesian bootstrap approximation thereof.

2.2 Gibbs posterior distributions

The Gibbs measure has its origins in statistical physics but a version of it has received attention in the statistics, machine learning, and econometrics literature; see, e.g., Bissiri et al. (2016), Zhang (2006a,b), and Chernozhukov and Hong (2003). Some recent statistical applications include data mining (Jiang and Tanner 2008), clinical trials (Syring and Martin 2017), image analysis (Syring and Martin 2020b), actuarial science (Syring et al. 2019), and classifier performance assessment (Wang and Martin 2020). Below we define the Gibbs posterior and some features that will be relevant in what follows.

Let $X^n = (X_1, \dots, X_n)$ be an iid sample from some distribution P . Suppose there is some functional $\theta = \theta(P)$ that we are interested in estimating and making inference about. By the way this problem has been stated, it should be clear that θ generally cannot be understood as a model parameter, so we cannot expect that there is a likelihood function that can be used to connect the data to the quantity of interest. Instead, the setup assumes that the functional is defined via an optimization problem. That is, there exists a function $\ell_\theta(x)$ such that the true value θ^* of $\theta(P)$ is the minimizer of the function $R(\theta) = P\ell_\theta$; here, note that, as is customary in the literature, we denote the quantity of interest and a generic value of it with the same symbol, θ , and distinguish the true value θ^* where necessary. The function ℓ_θ is called the *loss* and $R(\theta)$ the corresponding *risk*. Since we do not know P , we also do not know the risk, so inference on θ requires that we replace P with the observed data in some way. In particular, define the empirical risk as $R_n(\theta) = \mathbb{P}_n \ell_\theta$, where \mathbb{P}_n is the empirical distribution of the data X^n . The estimator $\hat{\theta}_n$ derived by minimizing $R_n(\theta)$ is often called an *M-estimator* (e.g., Huber 1981).

Empirical risk minimization is a common task in machine learning and can be challenging because the data-dependent objective function R_n is not always well-behaved. As an alternative to optimization, the *PAC-Bayes* literature (e.g., Alquier 2008; McAllester 1999)—where PAC stands for probability approximately correct—proposed to construct a distribution that concentrates on θ values for which $R_n(\theta)$ is small. That distribution is the Gibbs posterior and is given by

$$\Pi_n(B) = \frac{\int_B e^{-\omega n R_n(\theta)} \Pi(d\theta)}{\int_{\mathbb{R}^d} e^{-\omega n R_n(\theta)} \Pi(d\theta)}, \quad B \subseteq \mathbb{R}^d, \quad (4)$$

where Π is a prior distribution and $\omega > 0$ is called the *learning rate*. The choice to use

the empirical risk function in (4) is crucial. Here we adopt the perspective of Bissiri et al. (2016), who show that, when—like in our present application—the quantity of interest is defined as the minimizer of the expected loss, $R(\theta) = P\ell_\theta$, the *proper generalization of Bayesian inference*, in the sense of coherent updates of beliefs, leads to the formulation in (4) with the empirical version of the expected loss, $R_n(\theta) = \mathbb{P}_n\ell_\theta$. So, while other choices of “pseudo-posterior distributions” are possible, these lack justification as a proper generalization of Bayesian inference.

It should be emphasized that the introduction of the learning rate $\omega > 0$ in (4) is not an arbitrary choice being made by us, it is a technical artifact of the generalized Bayes posterior construction in, e.g., Bissiri et al. (2016). Both the prior and learning rate play crucial roles in determining the Gibbs posterior’s practical performance. Unfortunately, neither are fully determined by the context/data, so effort is required from the user.

- Of course, one can use a vague/flat prior, which we do in our simulation study below to avoid confounding the Gibbs posterior’s performance with the effects of using an informative prior, but this is not our recommendation. Since a multivariate quantile is a real-world quantity, and not the parameter of an artificial model, it is possible that genuine prior information is available and, in such cases, that prior information absolutely should be used.
- The learning rate is analogous to the tuning parameter in machine learning algorithms in the sense that there is no “true” ω that can be learned from the data. However, data-driven choices of the learning rate are still possible, and a number of such methods have been proposed in the recent literature, e.g., Grünwald (2012), Holmes and Walker (2017), Lyddon et al. (2019), and Syring and Martin (2019); a comparison of these can be found in Wu and Martin (2020). More details about learning rate selection are given in Section 3.3.

3 Gibbs posteriors for multivariate quantiles

3.1 Definition

Suppose we have an iid sample X_1, \dots, X_n from a distribution P on \mathbb{R}^d . Since the ℓ_1 -median $\theta(P)$ is the same as $Q_P(0)$, we will discuss the Gibbs posterior construction for a geometric quantile $Q(u)$ with ℓ_r -norm for some fixed $r \in (1, \infty)$ and fixed $u \in B_q^{(d)}$. For simplicity, we will denote $\hat{\theta}_n = \hat{Q}_n(u)$ and $\theta^* = Q_P(u)$ from now on.

Since the quantity of interest is the minimizer of a function defined by an expectation, in (3), it makes sense to define the loss ℓ_θ as that function inside the expectation. Precisely, we take $\ell_\theta(x) = \Phi_r(u, x - \theta)$, where, again, u and r are fixed. Then the risk is $R(\theta) = P\ell_\theta$, minimized at θ^* , and the empirical risk is

$$R_n(\theta) = \mathbb{P}_n\ell_\theta = \frac{1}{n} \sum_{i=1}^n \{\|X_i - \theta\|_r + \langle u, X_i - \theta \rangle\}, \quad (5)$$

minimized at $\hat{\theta}_n$. Given a prior distribution Π for θ , the Gibbs posterior distribution Π_n is defined like in (4). It follows from Hölder’s inequality that $R_n(\theta) \geq 0$, so if the prior is proper, then the denominator in (4) is finite and the Gibbs posterior is well-defined.

Therefore, Π_n is just an ordinary probability distribution, and features of that distribution can be extracted and summarized via the standard Markov chain Monte Carlo methods. For the theoretical analysis that follows, we assume that the learning rate ω is a fixed constant, but we will recommend a data-driving choice of ω in Section 3.3.

While the definition of the Gibbs posterior in (4) appears to be only a modest generalization of the familiar Bayesian definition, there are some important differences that deserve emphasis. First, in this multivariate quantile setting, θ is not a parameter indexing a statistical model, so *there is no likelihood function for θ* and, consequently, no direct Bayesian posterior distribution for θ . Therefore, any ordinary Bayesian approach to this problem would necessarily be indirect, i.e., define a statistical model with parameter, say, ψ , introduce a prior distribution for ψ , evaluate the corresponding posterior, and then marginalize to θ , which is now a function of ψ . So, despite the superficial similarity between the Gibbs and Bayesian posterior distributions, they are in fact quite different. Second, along similar lines, since θ is a real-world quantity that exists independently of a statistical model, there might be genuine prior information available about it. That the Gibbs posterior is direct implies that this prior information can be readily incorporated in (4). Compare this to an indirect Bayesian posterior distribution where the model parameter ψ often has limited real-world interpretation and, therefore, requires a non-informative prior, making it impossible to incorporate whatever prior information about θ might be available in a given application.

3.2 Asymptotic properties

First, we investigate the Gibbs posterior concentration rate, i.e., the radius of the smallest ball around θ^* to which the posterior asymptotically assigns all of its mass, as $n \rightarrow \infty$. **We focus on the case $d \geq 2$, and impose the following mild conditions on P .**

Assumption 1. *P admits a density p that is continuous and bounded away from 0 on a compact set $\mathbb{X} \subset \mathbb{R}^d$ containing θ^* and having non-empty interior.*

Assumption 2. *The density p is bounded away from ∞ on compact subsets of \mathbb{R}^d .*

An important consequence (see the proof of Lemma 1 in the Appendix) of Assumptions 1–2 is that the function R is twice differentiable at θ^* , where $\dot{R}(\theta^*) = 0$ and $V_{\theta^*} := \ddot{R}(\theta^*)$ is positive definite; here, dot and double-dot correspond to first and second derivatives with respect to θ , the gradient vector and the Hessian matrix, respectively.

Assumption 3. *The prior distribution Π has a density π which is continuous and bounded away from 0 in a neighborhood of θ^* .*

Theorem 1. *Under Assumptions 1–3, for any $\omega > 0$, the Gibbs posterior Π_n for the multivariate quantile satisfies*

$$P^n \Pi_n(\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2 > a_n n^{-1/2}\}) = o(1), \quad n \rightarrow \infty,$$

where $a_n \rightarrow \infty$ is any diverging sequence.

Proof. See Appendix A.2. □

The concentration rate result in Theorem 1 holds for all $\omega > 0$, that is, there are no restrictions on ω needed to achieve the target root- n rate. However, the limiting Gibbs posterior distribution shape in Theorem 2 below and, hence, the practical performance of the Gibbs posterior distribution, does depend on ω , so we recommend a data-driven choice as described in Section 3.3.

Next, we will prove a Bernstein–von Mises theorem for the Gibbs posterior, that is, the Gibbs posterior can be approximated by a Gaussian distribution in a total variation sense as $n \rightarrow \infty$. Before formally stating this result, we need a bit more notation. The loss $\theta \mapsto \ell_\theta(x)$ can be differentiated for P -almost all x , and the j^{th} component of the gradient vector, $\dot{\ell}_\theta(x)$, is given by

$$\dot{\ell}_\theta(x)_j = \frac{|x_j - \theta_j|^{r-1}}{\|x - \theta\|_r^{r-1}} \text{sign}(\theta_j - x_j) - u_j, \quad j = 1, \dots, d,$$

where, again, r and u are fixed, and $\text{sign}(\cdot)$ denotes the *signum* function. Now set $\Delta_{n,\theta^*} = n^{-1/2} \sum_{i=1}^n V_{\theta^*}^{-1} \dot{\ell}_{\theta^*}(X_i)$.

Theorem 2. *Under Assumptions 1–3, the sequence of centered and scaled Gibbs posteriors, with any learning rate $\omega > 0$, approaches a sequence of d -variate normal distributions in total variation, that is,*

$$\sup_B \left| \Pi_n(\{\theta : n^{1/2}(\theta - \theta^*) \in B\}) - \mathbf{N}_d(B \mid \omega \Delta_{n,\theta^*}, (\omega V_{\theta^*})^{-1}) \right| = o_P(1), \quad n \rightarrow \infty.$$

Proof. See Appendix A.3. □

A few technical remarks about the theorem and its proof are in order.

- The proof proceeds by checking the available sufficient conditions for Bernstein–von Mises theorems, e.g., in Chernozhukov and Hong (2003), Kleijn and van der Vaart (2012), etc. Here we opt to follow the latter reference whose results are more flexible and easier to apply in other similar applications.
- A similar result would hold in examples other than multivariate quantiles. The critical condition is that the empirical risk R_n satisfies a version of the *local asymptotic normality* condition, i.e., for every compact set $K \subset \mathbb{R}^d$,

$$\sup_{h \in K} \left| n\{R_n(\theta^* + hn^{-1/2}) - R_n(\theta^*)\} - h^\top V_{\theta^*} \Delta_{n,\theta^*} - \frac{1}{2} h^\top V_{\theta^*} h \right| = o_P(1). \quad (6)$$

So we can expect similar conclusions in any other problem for which (6) holds.

- Finally, like in van der Vaart (2000, p. 144), the normal approximation can be centered about an estimator that is asymptotically equivalent to Δ_{n,θ^*} . In particular, using the location shift invariance of the total variation distance, it follows from Theorem 2 that the Gibbs posterior Π_n is approximately $\mathbf{N}_d(\hat{\theta}_n, (\omega n V_{\theta^*})^{-1})$.

In a Bayesian setting, with a regular, well-specified model, a Bernstein–von Mises theorem ensures that inferences derived from the Bayesian posterior distribution are valid in a frequentist sense. For example, a $100(1 - \alpha)\%$ posterior credible set will have frequentist coverage probability approximately equal to $1 - \alpha$ for large n . The

reason for this Bayesian–frequentist connection is that the posterior distribution centers around, in that case, the maximum likelihood estimator, and the covariance matrix in the Bernstein–von Mises theorem is the inverse Fisher information matrix, which agrees with the asymptotic covariance matrix of the maximum likelihood estimator. However, when the model is misspecified, like in Kleijn and van der Vaart (2012), or, like here, where no model is specified at all, then this covariance matching is not guaranteed. Indeed, in our present case, the covariance matrix in the normal approximation to the Gibbs posterior is $(\omega V_{\theta^*})^{-1}$ whereas the asymptotic covariance matrix of $n^{1/2}(\hat{\theta}_n - \theta^*)$ is

$$\Gamma = V_{\theta^*}^{-1} P(\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^\top) V_{\theta^*}^{-1}, \quad (7)$$

which comes from the familiar sandwich formula. Since these two matrices are generally different, our Bernstein–von Mises theorem does not guarantee that inference drawn from the Gibbs posterior are valid in a frequentist sense. One way to avoid this covariance mismatch is to replace the empirical risk R_n in (4) with, e.g., a quadratic form like

$$\theta \mapsto \dot{R}_n(\theta)^\top \{ \mathbb{P}_n(\dot{\ell}_{\theta} \dot{\ell}_{\theta}^\top) \}^{-1} \dot{R}_n(\theta),$$

with the covariance mismatch correction term squeezed in. Another, following a suggestion in Yang et al. (2016), is to directly define a data-dependent distribution

$$\tilde{\Pi}_n(A) = \mathbb{N}_d(A \mid \hat{\theta}_n, n^{-1}\hat{\Gamma}), \quad (8)$$

where $\hat{\Gamma}$ is a suitable estimator of Γ as in (7). Both of these “pseudo-posterior distributions” have their merits, but they lack the interpretation of being proper generalized Bayes posteriors, which is our focus in this paper.

If we were in a traditional Bayesian setting and were unfortunate enough that our model was sufficiently misspecified that we get the aforementioned covariance matrix mismatch, then (a) we typically would not be aware of this problem and (b) there would be nothing we could do about it, aside from starting over with a different model. However, since we are working within a Gibbs framework, we are aware of and openly acknowledge that our posterior distribution is based on an effectively misspecified model and, moreover, we have a potential remedy: adjusting the learning rate.

It is easy to see that if $P(\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^\top) \propto V_{\theta^*}$, i.e., if the *generalized information equality* (Chernozhukov and Hong 2003) holds, then the covariance matrix in the normal approximation to the Gibbs posterior will be proportional to the asymptotic covariance matrix of the M-estimator $\hat{\theta}_n$. In that case, we can exactly correct for the covariance mismatch simply by tuning the learning rate. In general, however, simply tuning the scalar learning rate parameter cannot fully correct for the covariance mismatch, but it is still possible to find a learning rate such that credible sets derived from the Gibbs posterior have approximately the nominal frequentist coverage probability; see Section 3.3.

3.3 Choice of the learning rate

As we indicated in Section 2, the choice of learning rate is critical to the performance of methods derived from a Gibbs posterior distribution. This is especially important in our present situation because, as mentioned in the remarks following Theorem 2, the Gibbs posterior does not inherit the correct asymptotic shape. This covariance mismatch

Algorithm 1 — Gibbs Posterior Calibration (Syring and Martin 2019)

Fix a convergence tolerance $\epsilon > 0$ and an initial value $\omega^{(0)}$ of the learning rate ω . Take B bootstrap samples $\tilde{X}_1^n, \dots, \tilde{X}_B^n$ of size n . Set $t = 0$ and do the following.

1. Construct $100(1 - \alpha)\%$ credible set $C_{\omega^{(t)}, \alpha}(\tilde{X}_b^n)$ for every $b = 1, \dots, B$.
2. Evaluate the bootstrap estimate

$$\hat{c}_\alpha(\omega^{(t)}, \mathbb{P}_n) = \frac{1}{B} \sum_{b=1}^B \mathbb{1}\{C_{\omega^{(t)}, \alpha}(\tilde{X}_b^n) \ni \hat{\theta}_n\}$$

of the empirical coverage probability $c_\alpha(\omega^{(t)}, \mathbb{P}_n)$.

3. If $|\hat{c}_\alpha(\omega^{(t)}, \mathbb{P}_n) - (1 - \alpha)| < \epsilon$, then return $\omega^{(t)}$ as the output, else update $\omega^{(t)}$ to $\omega^{(t+1)}$ as

$$\omega^{(t+1)} = \omega^{(t)} + \kappa_t \{\hat{c}_\alpha(\omega^{(t)}, \mathbb{P}_n) - (1 - \alpha)\},$$

with $\kappa_t = (t + 1)^{-0.51}$, set $t \leftarrow t + 1$, and go back to Step 1.

is a common occurrence when a Bayesian model is misspecified but, unlike the traditional Bayesian setting where nothing can be done to overcome the misspecification bias, the Gibbs posterior has a learning rate that can be suitably chosen to correct for the mismatched asymptotic covariance matrix.

More specifically, following Syring and Martin (2019), if Π_n^ω denotes the Gibbs posterior with learning rate ω , then we aim to choose ω such that the frequentist coverage probabilities of credible sets derived from Π_n^ω are approximately equal to the nominal level. That is, for a desired significance level $\alpha \in (0, 1)$, if $C_{\omega, \alpha}(X^n)$ denotes a $100(1 - \alpha)\%$ credible set from the Gibbs posterior Π_n^ω , then the coverage probability is

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

i.e., the P -probability that $C_{\omega, \alpha}(X^n)$ contains $\theta(P)$. Of course, if P were known, then it would be possible to approximate the coverage probability using Monte Carlo and solve the equation, $c_\alpha(\omega; P) = 1 - \alpha$, using stochastic approximation (e.g., Robbins and Monro 1951). Since P is unknown in practice, Syring and Martin (2019) recommend a bootstrap version that replaces P with the empirical distribution, \mathbb{P}_n . We use their *Gibbs posterior calibration* algorithm (see Algorithm 1) for choosing the learning rate, which performs well in our experiments below.

For a quick visual illustration, consider a bivariate case, $d = 2$. Suppose we have $n = 100$ samples from a bivariate normal distribution as in Example 1 in Section 4.1. Using a relatively flat $\mathbf{N}_2(0, 10I_2)$ prior, and with the learning rate chosen according to Algorithm 1, samples from the corresponding Gibbs posterior distribution are shown in Figure 1(a). The same is shown in Figure 1(b), except where the data are sampled from a bivariate Laplace distribution as in Example 2 of Section 4.1. In addition to the Gibbs posterior samples, we also display the 95% credible region based on the normal approximation. That is, we first compute the posterior mean $\bar{\theta}$ and covariance matrix S based on the Monte Carlo samples, and then find the 95th percentile of the marginal

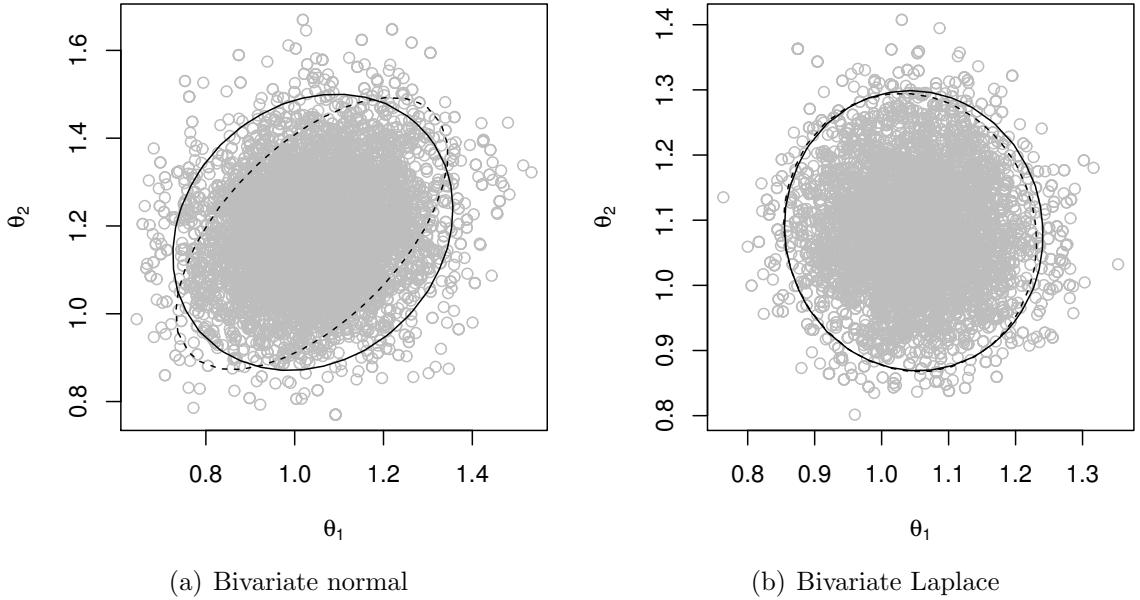


Figure 1: Gibbs posterior samples (gray), with learning rate chosen according to Algorithm 1, along with the 95% posterior credible region (solid) and the corresponding M-estimator confidence region (dashed).

posterior distribution for $\vartheta \mapsto (\vartheta - \bar{\theta})^\top S^{-1}(\vartheta - \bar{\theta})$, denoted by $r_{0.95}$. Then the 95% Gibbs posterior credible set is

$$\{\vartheta : (\vartheta - \bar{\theta})^\top S^{-1}(\vartheta - \bar{\theta}) \leq r_{0.95}\}.$$

Similarly, we compute the 95% confidence ellipse based on the asymptotic normality of the M-estimator/spatial median or, equivalently, the 95% credible set from (8), namely,

$$\{\vartheta : (\vartheta - \hat{\theta}_n)^\top \hat{\Gamma}_n(\vartheta - \hat{\theta}_n) \leq \chi^2_{2;0.95}\},$$

where $\chi^2_{2;0.95}$ is the 95th percentile of the chi-square distribution with 2 degrees of freedom. The boundaries of these two ellipses are overlaid on the plots of the Gibbs posterior samples. Clearly, in both cases, the contours of the Gibbs posterior are not of the same shape as the M-estimator confidence ellipse, a consequence of the covariance mismatch. However, by choosing the learning rate according to Algorithm 1, which is aiming to achieve the nominal 95% frequentist coverage rate, the Gibbs posterior credible ellipse is stretched to roughly match the confidence ellipse in the direction in which it is widest. And since the confidence ellipse achieves the nominal frequentist coverage probability, at least asymptotically, the Gibbs posterior credible ellipse will too. Of course, there is some loss of efficiency due to the covariance mismatch—which is the price one pays for a model-free posterior distribution—but, as the simulation results in Section 4.1 show, this loss of efficiency is not severe. In fact, in some cases, the Gibbs posterior credible regions are more efficient than those of other Bayesian methods.

4 Numerical results

4.1 Simulation study

In this section, we illustrate the finite sample performance of the Gibbs posterior of bivariate ℓ_1 -medians and quantiles, i.e., for $d = 2$. For the median, we would like to compare the Gibbs posterior's performance to that of both parametric and non-parametric Bayesian methods in situations when the components of the vector are correlated, and when the data has outliers. Aside from the d -variate normal distribution $\mathbf{N}_d(\mu, \Sigma)$ with mean vector μ and covariance matrix Σ , we also consider a d -variate Laplace distribution, denoted by $\mathbf{Lap}_d(\mu, \Sigma)$, with location vector μ and dispersion matrix Σ , with density for the standardized version, with $\mu = 0$ and $\Sigma = I_d$,

$$f(x) \propto \|x\|_2^{-(d-1)/2} e^{-2^{3/2}\|x\|_2}, \quad x \in \mathbb{R}^d.$$

Also, $\mathbf{Gamma}_d(s, r, V)$ denotes a d -variate gamma distribution with shape s , rate r , and correlation matrix V , constructed using a Gaussian copula (Xue-Kun Song 2000). The three specific examples we consider are as follows, each with sample size $n = 100$.

Example 1. $P = \mathbf{N}_2(\mu, \Sigma)$, where $\mu = (1, 1)^\top$, $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = 0.7$.

Example 2. $P = \mathbf{Lap}_2(\mu, \Sigma)$, where $\mu = (1, 1)^\top$ and $\Sigma = I_2$.

Example 3. $P = \mathbf{Gamma}_2(1, 1, V)$, where $V_{11} = V_{22} = 1$ and $V_{12} = 0.5$.

In each case, for the Gibbs posterior, we consider a bivariate normal prior for θ , namely, $\mathbf{N}_2((0, 0)^\top, 10I_2)$. We opt for a relatively non-informative prior, sacrificing one of the key benefits of the Gibbs posterior (Section 3.1) so that the performance of our Gibbs posterior is not confounded with the use of an informative prior. We also use a Metropolis–Hastings algorithm with transition kernel $Q(y \mid x) = \mathbf{N}_2(y \mid x, 0.01I_2)$ for drawing samples from the posterior distribution. The relevant summaries would be size and frequentist coverage of the 95% credible ellipses.

We compare the Gibbs posterior performance to that of parametric and non-parametric Bayesian methods. For the parametric Bayes model, we consider

$$(X_1, \dots, X_n) \mid \theta \stackrel{\text{iid}}{\sim} \mathbf{N}_2(\theta, \sigma^2 I_2) \quad \text{and} \quad \theta \sim \mathbf{N}_2((0, 0)^\top, 10I_2) \quad \sigma^{-2} \sim \mathbf{Gamma}(1, 1).$$

This is quite simple, and we can use a Gibbs sampler for posterior inference. While the model and corresponding analysis is simple, the concern is potential model misspecification bias. As a more robust alternative, one can consider a nonparametric Bayesian formulation. As suggested by Bhattacharya and Ghosal (2020), assume

$$(X_1, \dots, X_n) \mid P \stackrel{\text{iid}}{\sim} P \quad \text{and} \quad P \sim \mathbf{DP}(\alpha),$$

where $\mathbf{DP}(\alpha)$ denotes a Dirichlet process distribution with base or centering measure α (e.g., Ghosal and van der Vaart 2017, Ch. 4). Here we choose $\alpha = 2 \times \mathbf{N}_2((0, 0)^\top, I_2)$. It is well known that the Dirichlet process prior is conjugate, so the posterior distribution for P , given X^n , is also a Dirichlet process, which is relatively simple to work with. However, the quantity of interest is $\theta = \theta(P)$, a functional of P , so some non-trivial marginalization is required. Specifically, we sample P from the Dirichlet process posterior distribution,

Example	r	Gibbs	PBayes	NPBayes
1	2	0.950 (0.42)	0.900 (0.33)	0.975 (0.37)
	3	0.955 (0.40)	0.900 (0.33)	0.975 (0.38)
2	2	0.950 (0.21)	0.925 (0.34)	0.920 (0.19)
	3	0.960 (0.20)	0.925 (0.34)	0.970 (0.21)
3	2	0.950 (0.31)	0.926 (0.35)	0.950 (0.40)
	3	0.965 (0.21)	0.910 (0.35)	0.949 (0.27)

Table 1: Estimated coverage probability and mean size (in parentheses) of 95% credible ellipses of the ℓ_1 -median (with ℓ_2 and ℓ_3 norms) for the Gibbs, parametric Bayes, and nonparametric Bayes posterior distributions.

and then evaluate θ as the minimizer of $\xi \mapsto P\|X - \xi\|_r$. Bhattacharya and Ghosal (2020) establish that this nonparametric Bayes marginal posterior for θ is asymptotically equivalent to the Gaussian distribution in (8), i.e., no covariance mismatch, so must be relatively close to the best possible frequentist solution.

To compare the four methods described above, we consider a measure of size and also the frequentist coverage probability of the 95% posterior credible sets. Suppose that we have samples $\theta_1, \dots, \theta_M$ from any one of the three posterior distributions, where $M = 5000$ is the Monte Carlo sample size. The 95% credible set is given by

$$\{\vartheta : (\vartheta - \bar{\theta})^\top S^{-1}(\vartheta - \bar{\theta}) \leq r_{0.95}\},$$

where $\bar{\theta}$ and S are the Monte Carlo sample mean and covariance matrix, respectively. The coverage probability is defined as usual and, as a measure of the credible set's size, we use $|S|r_{0.95}^d$, with $|S|$ denoting the determinant of the matrix S . Table 1 summarizes the size and coverage probability over 2000 replications for each of the three examples. It can be seen that the Gibbs posterior performs well compared to the parametric and nonparametric Bayesian approaches in terms of both size and coverage. Indeed, the Gibbs posterior has at least the nominal 95% coverage in every scenario, while the parametric and nonparametric Bayesian credible sets occasionally miss the target coverage, especially the parametric solution. The coverage performance of the Gibbs credible regions may not be surprising, given that we tuned the learning rate to achieve the nominal coverage. However, it is interesting to see that, despite the covariance mismatch, the coverage guarantees do not come with any perceptible loss of efficiency. In fact, in some cases, especially in Example 3, the Gibbs posterior credible sets might be more efficient.

We also consider inference on a general u^{th} geometric quantile other than the ℓ_1 -median, for $u = (0.2, 0.3)$. While one can imagine the ℓ_1 -median as a type of location parameter, the general geometric quantile has no such interpretation. This implies that there is no genuine likelihood function for this unknown, which is precisely the case in which the Gibbs framework is most valuable. Here we compare the performance of the Gibbs posterior with the nonparametric Bayesian approach described above, again in terms of coverage probability and credible region size. The results are presented in Table 2. Note here that the Gibbs posterior achieves or exceeds the target 0.95 coverage probability while being more (or at least no less) efficient than the nonparametric Bayes solution. It is worth noting here that we have used a relatively flat prior for our illustra-

Example	Gibbs	NPBayes
1	0.990 (0.56)	0.955 (0.71)
2	0.950 (0.43)	0.925 (0.44)
3	0.950 (0.31)	0.945 (0.41)

Table 2: Estimated coverage probability and mean size (in parentheses) of 95% credible ellipses of $u = (0.2, 0.3)$ th geometric quantile (with ℓ_2 norm) for the Gibbs and nonparametric Bayes posterior distributions.

tion. But when relevant prior information about the quantile is available, this can readily be incorporated into the Gibbs posterior because it deals directly with the quantile, which would lead to additional gains in efficiency. The nonparametric Bayes formulation, on the other hand, focuses on the distribution P and, hence, works only indirectly with $\theta = \theta(P)$, so it is not at all clear how to incorporate prior information about θ to achieve those same efficiency gains.

4.2 Real data analysis

The Egyptian skulls dataset (Hand et al. 1994) consists of $d = 4$ measurements—namely, maximal breadth, basibregmatic height, basialveolar length, and nasal height—taken on $n = 150$ ancient Egyptian skulls from five time epochs; these data are available in the `HSAUR` package in R. The mean effect of time was removed by fitting a linear model and extracting the residuals, and we take these 4-dimensional residual vectors as our data X^n . A first thought would be to assume multivariate normality and carry out a standard analysis. However, formal tests of normality are conflicting: marginal tests of normality reject while tests of multivariate normality do not reject (e.g., Tokdar and Martin 2021). So, we proceed with the construction of a Gibbs posterior that does not require us to decide about normality in this difficult case.

For simplicity, here we will focus on the median of the 4-dimensional distribution, although other quantiles could be handled similarly. Since these data already have time trends removed, we expect that the distribution’s center should be roughly near the origin, so we take a normal prior with zero mean, but with covariance matrix $10I_4$. Since we are not at the $n \rightarrow \infty$ limit, we expect some non-elliptical shape in the Gibbs posterior, and the goal of this analysis is to investigate that shape. Figure 2 summarizes the marginal and pairwise Gibbs posterior distributions after scaling the learning rate according to Algorithm 1. As expected, even on these limited low-dimensional summaries, the Gibbs posterior does not appear to be exactly normal, but the results are quite reasonable. For comparison, we also show the marginal plug-in densities and 95% pairwise confidence ellipses based on the asymptotically normal sampling distribution of the M-estimator, the sample spatial median. Clearly, these margins of the Gibbs posterior are centered in roughly the correct place but, most importantly, and thanks to the calibration framework in Algorithm 1, the Gibbs posterior spread tends to be wider in some directions than that of the M-estimator sampling distribution. While some might view this wider spread as an indication of some inefficiency, we believe the wider spread is necessary for valid uncertainty quantification in finite samples, not just in the idealistic $n \rightarrow \infty$ case.

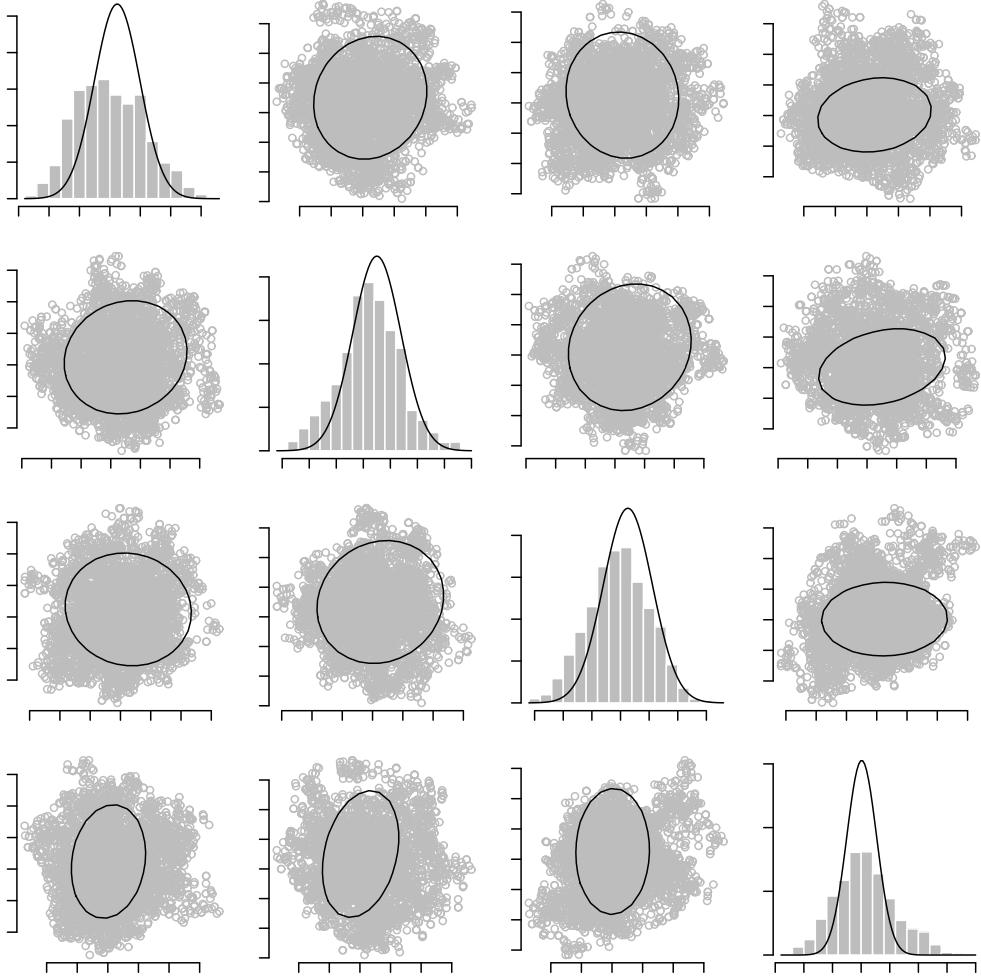


Figure 2: Marginal and pairwise Gibbs posterior distributions (gray) for the 4-dimensional spatial median of the Egyptian skull data described in Section 4.2. Overlaid (black) are approximate marginal sampling distributions and 95% pairwise confidence ellipses based on asymptotic normality of the M-estimator.

To compare the performance of our Gibbs posterior with a Bayes solution that assumes normality, we look at the posterior of the empirical risk evaluated on a held-out testing set. That is, we split the data into a training and testing set—the first 100 samples are training and the last 50 are testing—construct a posterior distribution for the median θ using the training data, then evaluate the corresponding posterior distribution for $R_{\text{test}}(\theta)$, where R_{test} is the empirical risk function in (5) based on the testing data set only. Figure 3 plots (kernel density estimates of) the posterior distribution of log empirical risk difference,

$$\log\{R_{\text{test}}(\theta) - \min_{\vartheta} R_{\text{test}}(\vartheta)\}, \quad (9)$$

based on the testing data. We follow the same Gibbs formulation as above; for the Bayes solution, we assume $P = \mathbf{N}_4(\theta, \Sigma)$ and use the conjugate normal-inverse Wishart prior for (θ, Σ) . The figure shows that the Gibbs posterior distribution of the log empirical risk difference is centered to left of that for Bayes, which is an indication that the former

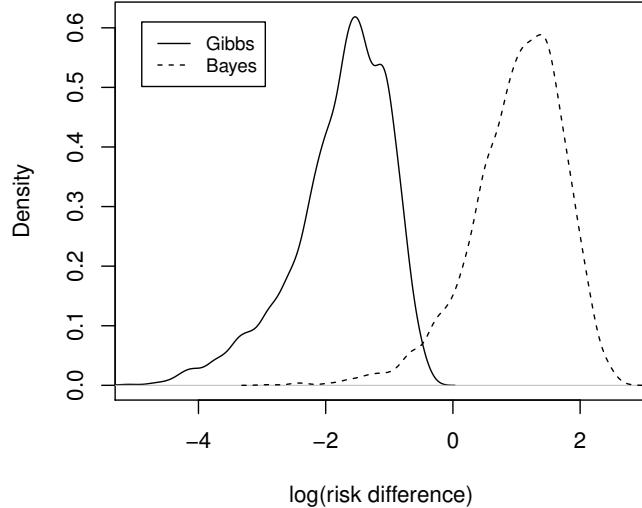


Figure 3: Plots of (kernel density estimates of) the Gibbs and Bayes posterior distribution for the log risk difference (9) based on the testing data.

is more concentrated around θ values that make the out-of-sample risk small than the latter. Since the training and testing data are not fundamentally different, this suggests that there is some bias created by the assumption of multivariate normality made by the parametric Bayesian solution. The Gibbs posterior, however, is apparently not susceptible to this model misspecification bias.

5 Concluding remarks

In this paper, we have studied multivariate medians and quantiles in a Gibbs posterior framework. Our approach does not need a model and is free from the potential issues that may arise in a model-based parametric Bayesian approach, in particular, model misspecification bias. The Gibbs posterior is simple to use and is also theoretically justified in the sense that the posterior concentrates around the true multivariate quantile at the optimal $n^{-1/2}$ rate, and has a Bernstein–von Mises property, i.e., it can be approximated by a suitable Gaussian distribution centered at the sample spatial median. We also pointed out that this Gaussian approximation holds in other problems, not just multivariate quantiles, provided that the empirical risk satisfies a version of the local asymptotic normality property.

A unique feature of the Gibbs posterior is its dependence on the choice of learning rate. On one hand, this dependence might seem like a disadvantage, since the learning rate is not determined by the context of the problem, and there is no universally accepted choice. On the other hand, as we argued here, being able to choose the learning rate is an advantage in the sense that it provides the flexibility necessary to at least partially correct for the covariance mismatch in the Gaussian approximation discussed in Section 3.2. Here we recommend the data-driven learning rate selection procedure of Syring and Martin (2019), summarized in Algorithm 1, as it aims to set ω so that the Gibbs posterior credible

region achieves the nominal frequentist coverage probability.

While the learning rate selection procedure described in Algorithm 1 works well empirically, there are still some opportunities for improvement and some unanswered questions. The main disadvantage of this strategy is having to do multiple Monte Carlo runs on each bootstrap sample; this could be improved by carrying out some of the steps in parallel. In terms of open questions, so far there is no theory to support the claim that choosing the learning rate according to Algorithm 1 will, as advertised, produce credible sets that achieve the nominal frequentist coverage. Syring and Martin (2019) argue that, under regularity conditions like those here that yield root- n convergence rates, there exists a value of ω such that the Gibbs credible regions achieve the nominal coverage. Then Algorithm 1 uses standard simulation-based techniques—bootstrap, Monte Carlo, and stochastic approximation—to find this solution. There is no reason to doubt that it would work, and the empirical results confirm this. Unfortunately, these three standard simulation-based techniques working in tandem make the algorithm quite difficult to analyze theoretically. That makes it an interesting open problem.

There are a couple of possible extensions of the work presented herein:

- It would be interesting to explore cases where the dimension d exceeds the sample size, i.e., a so-called “high-dimensional setting,” with $d \gg n$. For such cases, we would need to assume some low-dimensional structure in the high-dimensional θ^* , and then specify a prior distribution that would encourage this structure. Sparsity-inducing priors (e.g., Castillo and van der Vaart 2012; Martin and Ning 2020; van der Pas et al. 2017) have been popular in recent years, and one of the examples in Syring and Martin (2020a) shows that this kind of sparsity can be readily handled within the Gibbs framework, but the details for a sparse, high-dimensional multivariate quantile have yet to be worked out.
- The Gibbs posterior approach can also be used in multivariate quantile regression. Consider a linear regression set-up with a d -variate response vector y and a q -dimensional regressor x satisfying the linear model $y = \beta^\top x + e$, with β a $q \times d$ matrix of regression coefficients. For $u \in B_2^{(d)}$, and a sample (x_i, y_i) of response and regressor pairs, the u^{th} sample geometric quantile of y given x is obtained as

$$Q_{y|x}(u) = \arg \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \{ \|y_i - \beta^\top x_i\|_r + \langle u, y_i - \beta^\top x_i \rangle \}.$$

In a typical Bayesian approach, we would have to choose a model for the errors such that its quantile agrees with the target quantile; of course, there are many such models, so having to make such a choice puts the data analyst at risk of model misspecification bias. On the other hand, it is easy to formulate a Gibbs posterior framework, which is free of such modeling and the associated risks. We expect that the theoretical results for the Gibbs posterior presented herein would carry over to this more general setting, but we have yet to verify this conjecture.

Acknowledgments

Thanks go to two anonymous reviewers for their helpful comments that lead to a number of improvements to the manuscript, both in presentation and strength of results. This

work is partially supported by the U.S. National Science Foundation, DMS-1811802.

A Proofs

A.1 Preliminary results

Recall that $\ell_\theta(x) = \|x - \theta\|_r + \langle u, x - \theta \rangle$, and $R(\theta) = P\ell_\theta$, and the second derivative matrix is given by V_{θ^*} . First, we want to bound

$$\inf_{\|\theta - \theta^*\|_2 > \delta} R(\theta) - R(\theta^*). \quad (10)$$

Lemma 1. *Under Assumptions 1–2, there exists a constant $C > 0$ such that (10) is lower-bounded by $C\delta^2$ for all sufficiently small $\delta > 0$.*

Proof. The function R is twice-differentiable at θ^* , and the second derivative matrix is given by $V_{\theta^*} = \ddot{R}(\theta^*) = \int v_{\theta^*}(x) P(dx)$, where

$$v_{\theta^*}(x) = \frac{r-1}{\|x - \theta^*\|_r} \left[\text{diag} \left(\frac{|x_1 - \theta_1^*|^{r-2}}{\|x - \theta^*\|_r^{r-2}}, \dots, \frac{|x_d - \theta_d^*|^{r-2}}{\|x - \theta^*\|_r^{r-2}} \right) - \frac{yy^\top}{\|x - \theta^*\|_r^{2(r-1)}} \right],$$

with $y = y(x, \theta^*)$ being equal to

$$y = (|x_1 - \theta_1^*|^{r-1} \text{sign}(x_1 - \theta_1^*), \dots, |x_d - \theta_d^*|^{r-1} \text{sign}(x_d - \theta_d^*))^\top.$$

The existence of V_{θ^*} can be verified using Assumption 2, i.e., for a fixed $\theta \in \mathbb{R}^d$, with $d \geq 2$, if P has a density p that is bounded on compact subsets of \mathbb{R}^d , then the expectation of $\|X - \theta\|_r^{-1}$ is finite. **This statement can be proved by using a d -dimensional polar transformation, where the Jacobian would contain the $(d-1)^{\text{st}}$ power of the length of the radius vector (Bose and Chaudhuri 1993, Prop. 3.1).** Since the first derivative vanishes at θ^* , the Taylor expansion takes the form

$$R(\theta) - R(\theta^*) = \frac{1}{2}(\theta - \theta^*)^\top V_{\theta^*}(\theta - \theta^*) + o(\|\theta - \theta^*\|_2^2).$$

Since V_{θ^*} is positive definite, the proof follows by taking $C = \frac{1}{2}\lambda_{\min}(V_{\theta^*})$, where $\lambda_{\min}(M)$ returns the smallest eigenvalue of the matrix M . \square

Now, rewrite the Gibbs posterior distribution as

$$\Pi_n(A) = \frac{N_n(A)}{D_n} = \frac{\int_A e^{-\omega n\{R_n(\theta) - R_n(\theta^*)\}} \Pi(d\theta)}{\int e^{-\omega n\{R_n(\theta) - R_n(\theta^*)\}} \Pi(d\theta)}, \quad A \subseteq \mathbb{R}^d. \quad (11)$$

Below we will investigate the limiting behavior of $\Pi_n(A)$ for two kinds of sets A : the first is for a fixed $A = \mathbb{X}^c$, where \mathbb{X} is defined in Assumption 1, and the second is for a sequence of suitably shrinking sets A_n to be defined below.

Lemma 2. *Under Assumptions 1–2, for any $\omega > 0$, the Gibbs posterior Π_n satisfies $P^n \Pi_n(\mathbb{X}^c) = o(1)$ as $n \rightarrow \infty$.*

Proof. Without loss of generality, assume $\mathbb{X} = \{x : \|x - \theta^*\|_2 \leq K\}$ for $K > 0$. We start with the numerator $N_n(\mathbb{X}^c)$. By the law of large numbers, for any fixed unit vector u ,

$$R_n(\theta^* + Ku) - R_n(\theta^*) \rightarrow R(\theta^* + Ku) - R(\theta^*), \quad \text{in } P^n\text{-probability.}$$

Moreover, is easy to check that $\theta \mapsto R_n(\theta)$ is almost surely convex, so it follows from Lemma 1 in Hjort and Pollard (1993) that

$$\inf_u \{R_n(\theta^* + Ku) - R_n(\theta^*)\} \rightarrow \Delta \quad \text{in } P^n\text{-probability,}$$

where the infimum is over all unit vectors u , and $\Delta = \inf_u \{R(\theta^* + Ku) - R(\theta^*)\}$ is strictly positive. Then the event $\{N_n(\mathbb{X}^c) \leq e^{-\omega n(\Delta/2)}\}$ is implied by

$$\inf_u \{R_n(\theta^* + Ku) - R_n(\theta^*)\} > \Delta/2,$$

which has P^n -probability converging to 1.

For the denominator D_n , Lemma 4 below establishes that $D_n > \frac{1}{2}\varepsilon_n^d e^{-\omega(1+c_n)}$, with P^n -probability converging to 1, where c_n is such that $c_n \rightarrow \infty$ arbitrarily slowly. Then, with P^n -probability converging to 1,

$$\Pi_n(\mathbb{X}^c) = \frac{N_n(\mathbb{X}^c)}{D_n} \leq 2e^{-n\omega\Delta/2 + \omega(1+c_n) + d \log \varepsilon_n}.$$

We are free to choose $c_n = o(n)$, so the upper bound is vanishing. Therefore, we can conclude that $\Pi_n(\mathbb{X}^c) \rightarrow 0$ in probability. Since $\Pi_n(\mathbb{X}^c)$ bounded and converges to 0 in probability, the claim follows from the dominated convergence theorem. \square

Lemma 3. Define the empirical process $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n f - Pf)$ and, for the compact \mathbb{X} in Lemma 2, define the event

$$\mathcal{B}_n = \{X^n : |\mathbb{G}_n(\ell_\theta - \ell_{\theta^*})| > b_n \|\theta - \theta^*\|_2 \text{ for some } \theta \in \mathbb{X}\},$$

where $b_n > 0$ is any divergent sequence $b_n \rightarrow \infty$. Then $P^n(\mathcal{B}_n) = o(1)$ as $n \rightarrow \infty$.

Proof. Since the loss function is Lipschitz and the domain of $\theta \mapsto \ell_\theta - \ell_{\theta^*}$ is restricted to the compact \mathbb{X} , it follows from Example 19.7 in van der Vaart (2000) that $\{\ell_\theta - \ell_{\theta^*} : \theta \in \mathbb{X}\}$ is a Donsker class. This implies weak convergence and, in particular, stochastic equicontinuity of the empirical process $\theta \mapsto \mathbb{G}_n(\ell_\theta - \ell_{\theta^*})$ on \mathbb{X} . Therefore, the uniform-on- \mathbb{X} behavior of the empirical process is controlled by its behavior at finitely many points in \mathbb{X} . At any fixed $\theta \in \mathbb{X}$, Bernstein's inequality gives

$$P^n\{|\mathbb{G}_n(\ell_\theta - \ell_{\theta^*})| > t\} \leq 2 \exp\left\{-\frac{\frac{1}{2}t^2}{\|\theta - \theta^*\|_2^2 + t\|\theta - \theta^*\|_2}\right\}.$$

If we take $t = \{2b_n \|\theta - \theta^*\|_2^2\}^{1/2}$, then the right-hand side of the above display is

$$\exp\left\{-\frac{b_n \|\theta - \theta^*\|_2^2}{\|\theta - \theta^*\|_2^2 + (2b_n)^{1/2} \|\theta - \theta^*\|_2}\right\} \leq \exp\{-\text{constant} \times b_n^{1/2}\}.$$

Since $b_n \rightarrow \infty$, the upper bound is vanishing, so we get

$$P^n\{|\mathbb{G}_n(\ell_\theta - \ell_{\theta^*})| > b_n \|\theta - \theta^*\|_2\} = o(1), \quad \text{for each fixed } \theta \in \mathbb{X}.$$

Then this, stochastic equicontinuity, and the union bound gives the desired result. \square

Recall that the events of interest in Theorem 1 are given by $A_n = \{\theta : \|\theta - \theta^*\|_2 > a_n \varepsilon_n\}$, where $\varepsilon_n = n^{-1/2}$ and $a_n \rightarrow \infty$ is arbitrary. To prove that theorem, we need to show that $\Pi_n(A_n) \rightarrow 0$ in expectation. Our strategy is to find a lower bound on D_n and an upper bound on $N_n(A_n)$, both defined in (11), such that the ratio of these two bounds is vanishing. The next two lemmas accomplish each these two goals in turn.

Lemma 4. *Let $c_n > 0$ be any divergent sequence, and define the event*

$$\mathcal{C}_n = \{X^n : D_n \leq \frac{1}{2} \varepsilon_n^d e^{-\omega(1+c_n)}\}.$$

Then under the conditions of Theorem 1, $P^n(\mathcal{C}_n) = o(1)$ as $n \rightarrow \infty$.

Proof. Define the set $K_n = \{\theta : m(\theta) \vee v(\theta) \leq \varepsilon_n^2\}$, where $a \vee b = \max(a, b)$, and

$$m(\theta) = R(\theta) - R(\theta^*) \quad \text{and} \quad v(\theta) = P\{(\ell_\theta - \ell_{\theta^*}) - m(\theta)\}^2,$$

are the mean and variance of the loss difference, respectively. From the Lipschitz property of the loss and the Taylor approximation in the proof of Lemma 1, it follows that

$$\|\theta - \theta^*\|_2 \lesssim \varepsilon_n \implies m(\theta) \vee v(\theta) \leq \varepsilon_n^2.$$

Therefore, by Assumption 3,

$$\Pi(K_n) \geq \Pi(\{\theta : \|\theta - \theta^*\|_2 \lesssim \varepsilon_n\}) \gtrsim \varepsilon_n^d.$$

Next, define a standardized version of the empirical risk difference, i.e.,

$$Z_n(\theta) = \frac{n\{R_n(\theta) - R_n(\theta^*)\} - nm(\theta)}{\{nv(\theta)\}^{1/2}}.$$

This is a function of both θ and the data X^n , so define the upper level sets

$$\mathcal{Z}_n = \{(\theta, X^n) : |Z_n(\theta)| \geq c_n\},$$

where c_n is the sequence in the lemma statement. Also define the cross-sections

$$\mathcal{Z}_n(\theta) = \{X^n : (\theta, X^n) \in \mathcal{Z}_n\} \quad \text{and} \quad \mathcal{Z}_n(X^n) = \{\theta : (\theta, X^n) \in \mathcal{Z}_n\}.$$

Since we have

$$n\{R_n(\theta) - R_n(\theta^*)\} = nm(\theta) + \{nv(\theta)\}^{1/2} Z_n(\theta),$$

and m , v , and Z_n are suitably bounded on $K_n \cap \mathcal{Z}_n(X^n)^c$, we immediately get

$$D_n \geq \int_{K_n \cap \mathcal{Z}_n(X^n)^c} e^{-\omega nm(\theta) - \omega\{nv(\theta)\}^{1/2} Z_n(\theta)} \Pi(d\theta) \geq e^{-\omega(1+c_n)} \Pi\{K_n \cap \mathcal{Z}_n(X^n)^c\},$$

using the fact that $n\varepsilon_n^2 = 1$. From this lower bound, we get

$$\begin{aligned} P^n\{D_n \leq \frac{1}{2} \Pi(K_n) e^{-\omega(1+c_n)}\} &\leq P^n\left[e^{-\omega(1+c_n)} \Pi\{K_n \cap \mathcal{Z}_n(X^n)^c\} \leq \frac{1}{2} \Pi(K_n) e^{-\omega(1+c_n)}\right] \\ &= P^n\left[\Pi\{K_n \cap \mathcal{Z}_n(X^n)\} \geq \frac{1}{2} \Pi(K_n)\right] \\ &\leq \frac{2P^n \Pi\{K_n \cap \mathcal{Z}_n(X^n)\}}{\Pi(K_n)}, \end{aligned}$$

where the last line is by Markov's inequality. Now use Fubini's theorem:

$$\begin{aligned}
P^n \Pi \{ K_n \cap \mathcal{Z}_n(x^n) \} &= \int \int 1\{\theta \in K_n \cap \mathcal{Z}_n(x^n)\} \Pi(d\theta) P^n(dx^n) \\
&= \int \int 1\{\theta \in K_n\} 1\{\theta \in \mathcal{Z}_n(x^n)\} P^n(dx^n) \Pi(d\theta) \\
&= \int_{K_n} P^n \{\mathcal{Z}_n(\theta)\} \Pi(d\theta).
\end{aligned}$$

By Chebyshev's inequality, $P^n \{\mathcal{Z}_n(\theta)\} \leq c_n^{-2}$, and hence

$$P^n \{D_n \leq \frac{1}{2} \Pi(K_n) e^{-\omega(1+c_n)}\} \leq 2c_n^{-2}.$$

Putting everything together, since $\Pi(K_n) \gtrsim \varepsilon_n^d$, we have that

$$P^n(\mathcal{C}_n) = P^n(D_n \leq \frac{1}{2} \varepsilon_n^d e^{-\omega(1+c_n)}) \leq 2c_n^{-2} = o(1), \quad n \rightarrow \infty. \quad \square$$

Lemma 5. *Let \mathbb{X} and \mathcal{B}_n be as in Lemma 3, with the sequence b_n in \mathcal{B}_n such that $b_n = o(a_n)$. Then there exists a constant $k > 0$ such that*

$$P^n \{N_n(A_n \cap \mathbb{X}) 1(\mathcal{B}_n^c)\} \lesssim \varepsilon_n^d a_n^d e^{-\omega k a_n^2}, \quad \text{for all large } n.$$

Proof. For the empirical process \mathbb{G}_n defined above, write

$$R_n(\theta) - R_n(\theta^*) = R(\theta) - R(\theta^*) + n^{-1/2} \mathbb{G}_n(\ell_\theta - \ell_{\theta^*}).$$

Then the Gibbs posterior numerator at $A_n \cap \mathbb{X}$ can be decomposed as a sum of integrals over “shells” as follows:

$$\begin{aligned}
N_n(A_n \cap \mathbb{X}) &= \sum_{t=1}^{T_n} \int_{ta_n \varepsilon_n < \|\theta - \theta^*\|_2 < (t+1)a_n \varepsilon_n} e^{-\omega n \{R_n(\theta) - R_n(\theta^*)\}} \Pi(d\theta) \\
&\leq \sum_{t=1}^{T_n} e^{-C\omega t^2 a_n^2} \int_{\|\theta - \theta^*\|_2 < (t+1)a_n \varepsilon_n} e^{-\omega n^{1/2} \mathbb{G}_n(\ell_\theta - \ell_{\theta^*})} \Pi(d\theta),
\end{aligned}$$

where $C > 0$ is as in Lemma 1 and $T_n \rightarrow \infty$ is to account for the intersection with \mathbb{X} . The exponential term outside the integral results from the bound in Lemma 1 and the fact that $n\varepsilon_n^2 = 1$. On the event \mathcal{B}_n^c , the exponent in the integrand is bounded by $\omega b_n n^{1/2} \|\theta - \theta^*\|_2$, so

$$\begin{aligned}
N_n(A_n \cap \mathbb{X}) 1(\mathcal{B}_n^c) &\leq \sum_{t=1}^{T_n} e^{-C\omega t^2 a_n^2} e^{\omega(t+1)a_n b_n} \Pi(\{\theta : \|\theta - \theta^*\|_2 < (t+1)a_n \varepsilon_n\}) \\
&\lesssim (a_n \varepsilon_n)^d \sum_{t=1}^{\infty} e^{-\omega a_n^2 t^2 (C - 2b_n/a_n)} (t+1)^d,
\end{aligned}$$

where the last inequality uses the fact that the prior for θ has a bounded density on \mathbb{X} . Since $b_n \ll a_n$, the difference in the exponent will be bigger than some $k > 0$ for sufficiently large n . Therefore, the above sum is $\lesssim e^{-\omega k a_n^2}$, which proves the claim. \square

A.2 Proof of Theorem 1

To prove Theorem 1, we need to combine Lemmas 2–5. Towards this, write

$$\begin{aligned}
\Pi_n(A_n) &\leq \Pi_n(A_n \cap \mathbb{X}) + \Pi_n(\mathbb{X}^c) \\
&= \frac{N_n(A_n \cap \mathbb{X})}{D_n} + \Pi_n(\mathbb{X}^c) \\
&= \frac{N_n(A_n \cap \mathbb{X})}{D_n} \mathbf{1}(\mathcal{B}_n^c \cap \mathcal{C}_n^c) + \frac{N_n(A_n \cap \mathbb{X})}{D_n} \mathbf{1}(\mathcal{B}_n \cup \mathcal{C}_n) + \Pi_n(\mathbb{X}^c) \\
&\leq \frac{2N_n(A_n \cap \mathbb{X}) \mathbf{1}(\mathcal{B}_n^c)}{\varepsilon_n^d e^{-\omega(1+c_n)}} + \mathbf{1}(\mathcal{B}_n \cup \mathcal{C}_n) + \Pi_n(\mathbb{X}^c).
\end{aligned}$$

Taking expectation, we get

$$P^n \Pi_n(A_n) \leq \frac{2P^n N_n(A_n \cap \mathbb{X}) \mathbf{1}(\mathcal{B}_n^c)}{\varepsilon_n^d e^{-\omega(1+c_n)}} + P^n(\mathcal{B}_n) + P^n(\mathcal{C}_n) + P^n \Pi_n(\mathbb{X}^c).$$

The second, third, and fourth terms in the upper bound are $o(1)$ by Lemmas 3, 4, and 2, respectively. Lemma 5 says the first term in the upper bound above satisfies

$$\lesssim a_n^d e^{-\omega(ka_n^2 - c_n)}.$$

The sequence c_n is arbitrary, so we are free to take $c_n = o(a_n^2)$, in which case, the quantity in the above display vanishes, proving the claim.

A.3 Proof of Theorem 2

The proof begins by showing that $e^{-\omega n R_n(\theta)}$ satisfies a locally asymptotic normality condition, that is, for every compact set $K \subset \mathbb{R}^d$

$$\sup_{h \in K} |\log s_n(h) - \omega h^\top V_{\theta^*} \Delta_{n,\theta^*} - \frac{\omega}{2} h^\top V_{\theta^*} h| \rightarrow 0, \quad \text{in } P^n\text{-probability.} \quad (12)$$

where $s_n(h) = e^{-\omega n \{R_n(\theta^* + hn^{-1/2}) - R_n(\theta^*)\}}$. To show this,

$$\begin{aligned}
-\omega n \{R_n(\theta^* + hn^{-1/2}) - R_n(\theta^*)\} &= -\omega n \{\mathbb{P}_n(\ell_{\theta^* + hn^{-1/2}} - \ell_{\theta^*}) \\
&= -\omega n P(\ell_{\theta^* + hn^{-1/2}} - \ell_{\theta^*}) - \omega n^{1/2} \mathbb{G}_n(\ell_{\theta^* + hn^{-1/2}} - \ell_{\theta^*}).
\end{aligned}$$

Since the loss ℓ_θ is Lipschitz, it follows from Lemma 19.31 in van der Vaart (2000), that

$$\mathbb{G}_n \{\omega n^{1/2} (\ell_{\theta^* + hn^{-1/2}} - \ell_{\theta^*}) - \omega h^\top \dot{\ell}_{\theta^*}\} \rightarrow 0, \quad \text{in } P^n\text{-probability.} \quad (13)$$

Since R is twice differentiable at θ^* with second derivative matrix V_{θ^*} , (12) holds. Given that empirical risk difference has a suitable locally quadratic representation, it is intuitively clear that the Gibbs posterior distribution will take on a Gaussian shape. Confirming this intuition requires some care, but one can follow exactly the arguments used to prove Theorem 2.1 in Kleijn and van der Vaart (2012). Indeed, their setup concerns a sequence of statistical models with density functions $p_\theta^{(n)}$, and their misspecified Bayesian posterior distribution corresponds to our Gibbs posterior with the empirical risk replaced

by $R_n(\theta) = -n^{-1} \log p_\theta^{(n)}$ and, of course, $\omega = 1$. They show that if the $\omega = 1$ version (12) holds for their choice of R_n , and if the misspecified Bayes posterior concentrates around θ^* at the root- n rate, then the Bernstein–von Mises result holds. Our Gibbs formulation uses the empirical risk R_n , we have established (12) and the Gibbs posterior’s root- n rate from Theorem 1, so checking the Bernstein–von Mises property follows exactly the same steps as in Kleijn and van der Vaart.

References

Abdous, B. and Theodorescu, R. (1992). Note on the spatial quantile of a random vector. *Stat Probab Lett*, 13(4):333–336.

Alquier, P. (2008). PAC-Bayesian bounds for randomized empirical risk minimizers. *Math. Methods Statist.*, 17(4):279–304.

Babu, G. J. and Rao, C. R. (1989). Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population. *J. Multivar. Anal.*, 27(1):15–23.

Bhattacharya, I. and Ghosal, S. (2020). Bayesian inference on multivariate medians and quantiles. *Statist. Sinica*. To appear, http://www3.stat.sinica.edu.tw/ss_newpaper/SS-2020-0108_na.pdf.

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

Bose, A. and Chaudhuri, P. (1993). On the dispersion of multivariate median. *Ann. Inst. Statist. Math.*, 45(3):541–550.

Brown, B. (1983). Statistical uses of the spatial median. *J. R. Stat. Soc.*, 45(1):25–30.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

Chakraborty, B. (1999). On multivariate median regression. *Bernoulli*, 5(4):683–703.

Chakraborty, B. (2003). On multivariate quantile regression. *J. Statist. Plann. Inference*, 110(1-2):109–132.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Am. Stat. Assoc*, 91(434):862–872.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econom.*, 115(2):293–346.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, 26(3):299–316.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press.

Grünwald, P. (2012). The safe Bayesian: learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Comput. Sci.*, pages 169–183. Springer, Heidelberg.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.*, 12(4):1069–1103.

Hallin, M., Paindaveine, D., Šiman, M., Wei, Y., Serfling, R., Zuo, Y., Kong, L., and Mizera, I. (2010). Multivariate quantiles and multiple-output regression quantiles: From ℓ_1 optimization to halfspace depth (with discussion and rejoinder). *Ann. Statist.*, 38:635–703.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994). *A Handbook of Small Datasets*. Chapman and Hall/CRC, London.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. Unpublished manuscript, <http://www.stat.yale.edu/~pollard/Papers/convex.pdf>.

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.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York.

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

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381.

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478.

Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *Internat. J. Approx. Reason.*, 113:39–73.

Martin, R. and Ning, B. (2020). Empirical priors and coverage of posterior credible sets in a sparse normal mean model. *Sankhya A*, 82:477–498.

McAllester, D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, pages 164–170. ACM, New York.

Möttönen, J., Nordhausen, K., and Oja, H. (2010). Asymptotic theory of the spatial median. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, pages 182–193. Institute of Mathematical Statistics.

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

Small, C. G. (1990). A survey of multidimensional medians. *Int. Stat. Rev.*, 58(3):263–277.

Syring, N., Hong, L., and Martin, R. (2019). Gibbs posterior inference on value-at-risk. *Scand. Actuar. J.*, pages 1–10.

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

Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Syring, N. and Martin, R. (2020a). Gibbs posterior concentration rates under sub-exponential type losses. [arXiv:2012.04505](https://arxiv.org/abs/2012.04505).

Syring, N. and Martin, R. (2020b). Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image. *Ann. Statist.*, 48(3):1498–1513.

Tokdar, S. T. and Martin, R. (2021). Bayesian test of normality versus a Dirichlet process mixture alternative. *Sankhya B*, 83(1):66–96.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion and rejoinder). *Bayesian Anal.*, 12(4):1221–1274.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Wang, Z. and Martin, R. (2020). Model-free posterior inference on the area under the receiver operating characteristic curve. *J. Stat. Plan. Infer.*, 209:174–186.

Wu, P.-S. and Martin, R. (2020). A comparison of learning rate selection methods in generalized Bayesian inference. [arXiv:2012.11349](https://arxiv.org/abs/2012.11349).

Xue-Kun Song, P. (2000). Multivariate dispersion models generated from Gaussian copula. *Scand. J. Stat.*, 27(2):305–320.

Yang, Y., Wang, H. J., and He, X. (2016). Posterior inference in bayesian quantile regression with asymmetric laplace likelihood. *Int. Stat. Rev.*, 84(3):327–344.

Zhang, T. (2006a). From ϵ -entropy to KL-entropy: analysis of minimum information complexity density estimation. *Ann. Statist.*, 34(5):2180–2210.

Zhang, T. (2006b). Information theoretical upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory*, 52(4):1307–1321.