Robust Bayesian inference via coarsening

Jeffrey W. Miller*
Department of Biostatistics, Harvard University
and
David B. Dunson
Department of Statistical Science, Duke University

December 8, 2017

Abstract

The standard approach to Bayesian inference is based on the assumption that the distribution of the data belongs to the chosen model class. However, even a small violation of this assumption can have a large impact on the outcome of a Bayesian procedure. We introduce a novel approach to Bayesian inference that improves robustness to small departures from the model: rather than conditioning on the event that the observed data are generated by the model, one conditions on the event that the model generates data close to the observed data, in a distributional sense. When closeness is defined in terms of relative entropy, the resulting "coarsened" posterior can be approximated by simply tempering the likelihood—that is, by raising the likelihood to a fractional power—thus, inference can usually be implemented via standard algorithms, and one can even obtain analytical solutions when using conjugate priors. Some theoretical properties are derived, and we illustrate the approach with real and simulated data using mixture models, autoregressive models of unknown order, and variable selection in linear regression.

Keywords: Model error, model misspecification, power likelihood, relative entropy, tempering, Wasserstein distance.

*The authors gratefully acknowledge support from the National Science Foundation (NSF) grant DMS-1045153 and the National Institutes of Health (NIH) grants R01ES020619 and 5R01ES017436.

1 Introduction

In most applications, statistical models are idealizations that are known to provide only an approximation to the distribution of the observed data. One might hope that departures from the model, if sufficiently small, would not significantly impact inferences. Often this does seem to be the case, but sometimes inferences are sensitive to small perturbations away from the assumed model, especially if the sample size is large. This article focuses on the problem of defining alternatives to the usual likelihood function that are designed to be robust to a small amount of mismatch between the assumed model and the distribution of the observed data. Although the concepts are general, we concentrate on Bayesian approaches, using our modified likelihoods in place of the usual likelihood. We are focused on robustness to the form of the likelihood, in contrast to most previous work on robust Bayes which focuses on robustness to the choice of prior.

Instead of using the standard posterior obtained by conditioning on the event that the observed data are generated by the model—which is incorrect when there is a perturbation—our approach is to condition on the event that the empirical distribution of the observed data is close to the empirical distribution of data generated by the model, with respect to some discrepancy between probability measures. We refer to this as a coarsened posterior, or c-posterior, for short. This corresponds to using a modified likelihood.

One can control the type of robustness exhibited by a c-posterior via the choice of discrepancy. For instance, robustness to outliers can be obtained by using a discrepancy that is not strongly affected by moving a small amount of probability mass to an outlying region (e.g., 1st Wasserstein distance). Alternatively, robustness to slight changes in the shape of the distribution—which is our primary interest in this paper—can be obtained by using a discrepancy that is tolerant of such changes, such as relative entropy.

It works out particularly well to use relative entropy (i.e., Kullback–Leibler divergence), since in this case the c-posterior can be approximated by the "power posterior" obtained by simply raising the likelihood to a certain fractional power. Consequently, one can usu-

ally do approximate inference using standard algorithms with no additional computational burden—in fact, the mixing time of Markov chain Monte Carlo (MCMC) samplers will typically be improved, since the likelihood is tempered. Further, when using exponential families and conjugate priors, one can even obtain analytical expressions for quantities such as a robustified marginal likelihood.

The main novel contributions of the paper are: (1) introducing the idea of the c-posterior, (2) providing a calibration method for choosing an appropriate amount of coarsening, (3) empirically demonstrating how the c-posterior can easily be used to perform robust inference in a wide variety of examples, using real and simulated data, (4) establishing the asymptotic form of the c-posterior when certain limits are taken, (5) proving that the c-posterior exhibits robustness to small perturbations from the assumed model (that is, robustness to the form of the likelihood), and (6) proving that the power posterior is a good approximation to the relative entropy c-posterior when n is either large or small relative to the amount of coarsening.

The paper is organized as follows. Section 2 describes the coarsening approach. Section 3 illustrates the approach on a toy Bernoulli example, and Section 4 introduces a technique for choosing an appropriate amount of coarsening. In Sections 5–8, we apply the coarsening approach to (§5) mixture models and clustering, (§6) autoregressive models of unknown order, (§7) variable selection in linear regression, and (§8) a toy Normal example using Wasserstein distance. The supplementary material contains further discussion, theoretical results, and additional details.

2 Method

For now, we assume an i.i.d. setting, but the approach generalizes to time series and regression (see Sections 6 and 7). Suppose we have a model $\{P_\theta : \theta \in \Theta\}$ along with a prior Π on Θ, and suppose there is a point $\theta_I \in \Theta$ representing the parameters of the *idealized distribution* of the data. The interpretation is that θ_I is the true state of nature about

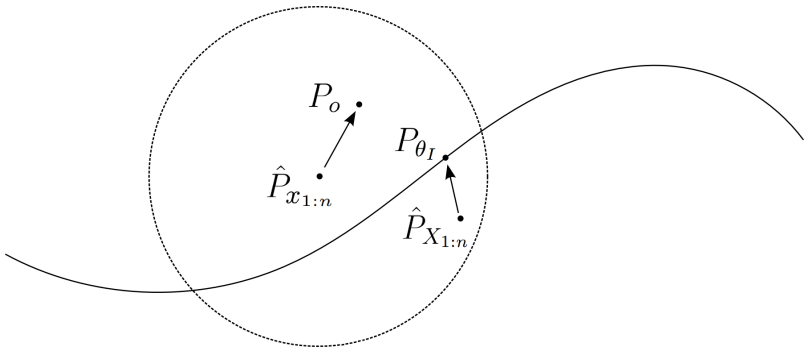


Figure 1: Notional schematic diagram of the idea behind the c-posterior. The ambient space is the set of probability distributions on $\mathcal{X}$, and the curve represents the subset of distributions in the parametrized family $\{P_\theta : \theta \in \Theta\}$. The idealized distribution P_{θ_I} is a point in this subset, and the empirical distribution $\hat{P}_{X_{1:n}}$ of the idealized data converges to P_{θ_I} as $n \to \infty$. Although $\hat{P}_{X_{1:n}}$ is not observed, it is known to be within an r-neighborhood of the empirical distribution $\hat{P}_{x_{1:n}}$ of the observed data, which, in turn, converges to the observed data distribution, P_o. The basic idea of the c-posterior approach is to condition on the event that $\hat{P}_{X_{1:n}}$ is within this neighborhood.

which one is interested in making inferences. Suppose there are some unobserved *idealized data* $X_1, \ldots, X_n \in \mathcal{X}$ that are i.i.d. from P_{θ_I}, and the *observed data* $x_1, \ldots, x_n \in \mathcal{X}$ are a perturbed version of $X_1, \ldots, X_n$ in the sense that $d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < r$ for some discrepancy $d(\cdot, \cdot)$ and some $r > 0$, where $\hat{P}_{x_{1:n}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ denotes the empirical distribution of $x_{1:n} = (x_1, \ldots, x_n)$. Suppose $x_1, \ldots, x_n$ behave like i.i.d. samples from some P_o, which we view as a perturbation of P_{θ_I}. For intuition, consider the diagram in Figure 1.

If there was no perturbation, then we would simply use the standard posterior—that is, we would condition on the event that $X_{1:n} = x_{1:n}$—however, when there is a perturbation, using the standard posterior is incorrect. If there is a known, easy-to-model process by which $x_{1:n}$ is generated from $X_{1:n}$, then we would simply augment the model to include this process—however, this process is often unknown or highly complex.

An alternative is to condition on the event that $d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < r$. In other words, rather than the standard posterior $\pi(\theta \mid X_{1:n} = x_{1:n})$, consider $\pi\big(\theta \mid d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < r\big)$. Since usually one will not have sufficient *a priori* knowledge to choose r, it makes sense to

put a prior on it, say $R \sim H$, independently of θ and $X_{1:n}$. Generalizing further, take a sequence of functions d_n such that $d_n(X_{1:n}, x_{1:n}) \geq 0$ is some measure of the discrepancy between $X_{1:n}$ and $x_{1:n}$.

Definition 2.1. We refer to $\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$ as a *c-posterior*.

To clarify the notation: if the prior Π has density π (with respect to some measure), then the c-posterior has density $\pi(\theta \mid Z = 1) \propto \pi(\theta)\,\mathbb{P}(Z = 1 \mid \theta)$ where $Z = \mathbb{1}(d_n(X_{1:n}, x_{1:n}) < R)$. In these expressions, $x_{1:n}$ is considered to be fixed, while $X_{1:n}$ and R are random variables; thus, the c-posterior is a function of $x_{1:n}$, but not $X_{1:n}$ and R since they are integrated out. (We use $\mathbb{1}(\cdot)$ to denote the indicator function: $\mathbb{1}(E) = 1$ if E is true, and $\mathbb{1}(E) = 0$ otherwise.) One can write the c-posterior as

$$\begin{aligned}\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) &\propto \pi(\theta)\,\mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \theta\big) \\ &= \pi(\theta) \int_{\mathcal{X}^n} G(d_n(x'_{1:n}, x_{1:n}))P_\theta^n(dx'_{1:n}) \qquad (2.1)\end{aligned}$$

where $G(r) = \mathbb{P}(R > r)$ and $\propto$ indicates proportionality with respect to θ. The intuitive interpretation is that, to use a rough analogy, this integral is like a convolution of P_θ^n (the distribution of $X_{1:n}$ given θ) with the "kernel" $G(d_n(X_{1:n}, x_{1:n}))$. The factor $\mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \theta\big)$ can be interpreted as a coarsened likelihood, or c-likelihood, however, it does not necessarily correspond to a probability distribution on $x_{1:n}$ given θ. The c-posterior should not be interpreted as implying a model for $x_{1:n}$ given θ; indeed, a key advantage of the method is that it allows one to avoid explicitly specifying a robust model.

In Supplement S3.1, we derive the form of the c-posterior as $n \to \infty$. Meanwhile, in Supplement S3.2, we show that under certain conditions, when n is fixed and the distribution of R converges to 0, the c-posterior converges to the standard posterior. In Supplement S3.3, we show that the c-posterior is robust to changes in P_o that are small with respect to the chosen discrepancy $d(\cdot, \cdot)$. There are different types of robustness that may be desired, and the type of robustness exhibited by the c-posterior can be customized through the choice of $d(\cdot, \cdot)$. A few potential candidates for $d(\cdot, \cdot)$ would be Kolmogorov–Smirnov (in the univariate setting), Wasserstein, or a maximum mean discrepancy (Gretton et al., 2006). When

P_θ and P_o have densities with respect to a common measure, it is also possible to accommodate discrepancies between densities such as relative entropy, Hellinger distance, and various divergences—even though they may be undefined for empirical distributions—by choosing $d_n(X_{1:n}, x_{1:n})$ to be a consistent estimator of $d(P_\theta, P_o)$.

In the examples, we focus mainly on relative entropy and variations thereof as our choice of $d(\cdot, \cdot)$, due to several appealing properties. In particular, there is an approximation that makes it unnecessary to explicitly compute $d_n(X_{1:n}, x_{1:n})$. We discuss this next.

2.1 Relative entropy c-posteriors

Suppose P_o and P_θ (for all $\theta \in \Theta$) have densities p_o and p_θ, respectively, with respect to some sigma-finite measure λ (e.g., Lebesgue measure, or counting measure on a discrete space). Define $d(P_\theta, P_o)$ to be the relative entropy, also known as Kullback–Leibler divergence,

$$d(P_\theta, P_o) = D(p_o \| p_\theta) = \int p_o(x) \Big(\log \frac{p_o(x)}{p_\theta(x)} \Big) \lambda(dx).$$

Suppose $d_n(X_{1:n}, x_{1:n})$ is a consistent estimator of $D(p_o \| p_\theta)$, and $R \sim \mathrm{Exp}(\alpha)$. Then one obtains the following approximation to the relative entropy c-posterior:

$$\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) \;\approx\propto\; \pi(\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n}, \tag{2.2}$$

where $\approx\propto$ means "approximately proportional to", i.e., the distribution on the left is approximately equal to the distribution proportional to the expression on the right, and

$$\zeta_n = \frac{1/n}{1/n + 1/\alpha} = \frac{\alpha}{\alpha + n}. \tag{2.3}$$

The approximation in Equation 2.2 is good when either $n \gg \alpha$ or $n \ll \alpha$ (Corollary S3.4, Theorem S3.6), under mild conditions. Empirically we find that the approximation can be quite accurate (see Figure 2). It makes intuitive sense that the approximation would be good in both the large-sample ($n \gg \alpha$) and small-sample ($n \ll \alpha$) regimes, when one considers the convolution representation in Equation 2.1. Also, note that $\zeta_n \approx \alpha/n$ when $n \gg \alpha$, whereas $\zeta_n \approx 1$ when $n \ll \alpha$, and ζ_n smoothly interpolates between these two

regimes. For the motivation behind the particular form of the power ζ_n, see Supplement S5. Section 4 introduces a technique for choosing α in a data-driven way.

A key feature of Equation 2.2 is that it enables one to approximate the c-posterior without explicitly computing the relative entropy estimates $d_n(X_{1:n}, x_{1:n})$, which would normally involve computing a density estimate of p_o in order to handle the entropy term $-\int p_o \log p_o$ in $D(p_o \| p_\theta)$. Since this entropy term is constant with respect to θ, it is absorbed into the constant of proportionality. Using an $\text{Exp}(\alpha)$ prior on R is not important for robustness (indeed, our theoretical results in Supplement S3 allow a very large class of distributions on R); choosing $R \sim \text{Exp}(\alpha)$ is only important for obtaining a computationally simple formula via cancellation of the entropy term.

Definition 2.2. Given $\zeta \in [0, 1]$, we refer to $\prod_{i=1}^n p_\theta(x_i)^\zeta$ as a *power likelihood*, and we refer to the distribution proportional to $\pi(\theta) \prod_{i=1}^n p_\theta(x_i)^\zeta$ as a *power posterior*.

Like the c-likelihood, the power likelihood should not be interpreted as implying a probability distribution on $x_{1:n}$ given θ. It should only be interpreted as an approximation to the c-likelihood, up to a constant of proportionality with respect to θ; see Equation S5.1. A useful interpretation of the power posterior is that it corresponds to adjusting the sample size from n to $n\zeta$, in the sense that the posterior will only be as concentrated as it would be if there were $n\zeta$ samples.

Due to its simple form, inference using the power posterior is often easy, or at least, no harder than inference using the ordinary posterior. We discuss two commonly occurring cases: analytical solution in the case of exponential families with conjugate priors, and Gibbs sampling in the case of conditionally conjugate priors.

2.1.1 Power posterior with conjugate priors

When using exponential families with conjugate priors, one can often obtain analytical expressions for integrals with respect to the power posterior. Suppose $p_\theta(x) = \exp\big(\theta^\intercal s(x) - \kappa(\theta)\big)$, where $s(x) = (s_1(x), \ldots, s_k(x))^\intercal$ are the sufficient statistics, and suppose $\pi(\theta) =$

$\pi_{\xi,\nu}(\theta)$ where $\pi_{\xi,\nu}(\theta) = \exp\big(\theta^{\intercal}\xi - \nu\kappa(\theta) - \psi(\xi,\nu)\big)$, noting that this defines a conjugate family. Then the power posterior is proportional to

$$\pi_{\xi,\nu}(\theta)\prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n} \propto \exp\Big(\theta^{\intercal}\big(\xi + \zeta_n\textstyle\sum_i s(x_i)\big) - (\nu + n\zeta_n)\kappa(\theta)\Big) \propto \pi_{\xi_n,\nu_n}(\theta), \tag{2.4}$$

where $\xi_n = \xi + \zeta_n \sum_i s(x_i)$ and $\nu_n = \nu + n\zeta_n$, and thus, the power posterior remains in the conjugate family.

For most conjugate families used in practice, simple analytical expressions are available for the log-normalization constant $\psi(\xi,\nu)$ as well as for many integrals with respect to $\pi_{\xi,\nu}(\theta)$. This enables one to obtain analytical expressions for many quantities of inferential interest under the power posterior, thus providing approximations to the corresponding quantities under the relative entropy c-posterior. For instance, one obtains a marginal power likelihood, $\int_\Theta \pi_{\xi,\nu}(\theta)\prod_{i=1}^n p_\theta(x_i)^{\zeta_n} d\theta = \exp\big(\psi(\xi_n,\nu_n) - \psi(\xi,\nu)\big)$, which can be used to compute robustified Bayes factors and posterior model probabilities. Such c-posterior summaries are robust to perturbations to P_o that are small with respect to relative entropy, whereas usual Bayes factors and model probabilities can be very sensitive to such perturbations for large n (Supplement S4). In Section 3, we illustrate this approach in a toy example involving Bernoulli trials, and in Section 6, we use this approach to perform robust inference for the order of an autoregressive model.

2.1.2 MCMC on the power posterior

Often, it is desirable to place conditionally conjugate priors on the parameters—for instance, placing independent normal and inverse-Wishart priors on the mean and covariance of a normal distribution. In such cases, one can easily use Gibbs sampling on the power posterior, because for each parameter given the others, we are back in the case of a conjugate prior, and thus the full conditionals belong to the conjugate family (just as in Equation 2.4). In Section 5, we use Gibbs sampling for robust inference in mixture models by employing a conditional power posterior. In Section 7, we use Gibbs sampling for robust variable selection in linear regression with the power posterior. More generally, samples

can be drawn from the power posterior by using Metropolis–Hastings MCMC, with the power likelihood in place of the usual likelihood.

The mixing performance of MCMC with the power posterior will often be better than with the standard posterior, since raising the likelihood to a fractional power (i.e., a power between 0 and 1) has the effect of flattening it, enabling the sampler to more easily move through the space, particularly when there are multiple modes and n is large. Indeed, raising the likelihood to a fractional power—also known as tempering—is sometimes done in more complex MCMC schemes in order to improve mixing.

3 Toy example: Perturbed Bernoulli trials

The purpose of this toy example is to illustrate the method in the simplest possible setting, and to assess the accuracy of the power posterior approximation in a situation where the exact c-posterior can be computed easily. Suppose $X_1, \ldots, X_n$ i.i.d. $\sim$ Bernoulli(θ) represent the outcomes of n replicates of a laboratory experiment, and the team of experimenters is interested in testing $\mathrm{H}_0 : \theta = 1/2$ versus $\mathrm{H}_1 : \theta \neq 1/2$. The standard Bayesian approach is to define a prior probability for each hypothesis, say, $\Pi(\mathrm{H}_0) = \Pi(\mathrm{H}_1) = 1/2$, and define a prior density for θ in the case of H_1, say, $\theta|\mathrm{H}_1 \sim \mathrm{Uniform}(0, 1)$. Inference then proceeds based on the posterior probabilities of the hypotheses, $\Pi(\mathrm{H}_0|x_{1:n})$ and $\Pi(\mathrm{H}_1|x_{1:n}) = 1 - \Pi(\mathrm{H}_0|x_{1:n})$, where $x_{1:n} = (x_1, \ldots, x_n)$. If the observed data $x_1, \ldots, x_n$ are sampled i.i.d. from Bernoulli(θ), then the posterior is guaranteed to converge to the correct answer, that is, $\Pi(\mathrm{H}_0|x_{1:n}) \xrightarrow{\text{a.s.}} \mathbb{1}(\theta = 1/2)$ as $n \to \infty$.

In reality, however, it is likely that the observed data do not exactly follow the assumed model. For instance, some of the experiments may have been conducted under slightly different conditions than others (such as at different times or by different researchers), or some of the outcomes may be corrupted due to human error in carrying out the experiment.

A natural choice of discrepancy is the relative entropy between the empirical distributions of $x_{1:n}$ and $X_{1:n}$, $D(\hat{p}_x \| \hat{p}_X) = \sum_{i=0}^{1} \hat{p}_x(i) \log(\hat{p}_x(i)/\hat{p}_X(i))$, where $\hat{p}_x(1) = \bar{x}$ and

$\hat{p}_x(0) = 1 - \bar{x}$ in this example. This leads us to consider the following coarsened posterior for inferences about H_0 and H_1:

$$\Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x||\hat{p}_X) < R\big), \tag{3.1}$$

where $R \sim \mathrm{Exp}(\alpha)$. How should we choose α? If we have no *a priori* knowledge of the size of perturbation to expect, then we can use the calibration curve technique in Section 4. Otherwise, in this example, we can interpret the neighborhood size r in terms of Euclidean distance via the chi-squared approximation to relative entropy, $D(p\|q) \approx \frac{1}{2}\chi^2(p,q)$ (see Prop. S5.1). In particular, when $\bar{X} \approx 1/2$ we have $D(\hat{p}_x||\hat{p}_X) \approx 2|\bar{x} - \bar{X}|^2$. Thus, if we expect that the perturbation will shift the sample mean by no more than ε or so when $\mathrm{H}_0 : \theta = 1/2$ is true, then it makes sense to choose α so that $\mathbb{E}R \approx 2\varepsilon^2$. Since $\mathbb{E}R = 1/\alpha$, this suggests using $\alpha = 1/(2\varepsilon^2)$.

In this toy example, the c-posterior in Equation 3.1 can be computed exactly (see Supplement S7.1), however, in more complex cases, an approximation is needed. The power likelihood approximation from Section 2.1, when applied to this example, yields

$$\Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x||\hat{p}_X) < R\big) \approx 1/\big(1 + 2^{n\zeta_n} B(1 + n\zeta_n\bar{x},\, 1 + n\zeta_n(1 - \bar{x}))\big) \tag{3.2}$$

where $\zeta_n = \alpha/(\alpha + n)$ and $B(a, b)$ is the beta function (Supplement S7.1). Comparing this to the standard posterior,

$$\Pi\big(\mathrm{H}_0 \,\big|\, X_{1:n} = x_{1:n}\big) = 1/\big(1 + 2^n B(1 + n\bar{x},\, 1 + n(1 - \bar{x}))\big), \tag{3.3}$$

note that the only difference is that n has been replaced by $n\zeta_n$.

To illustrate numerically, suppose we would like to be robust to perturbations affecting $\bar{x}$ by roughly $\varepsilon = 0.02$ when H_0 is true. As described above, this corresponds to $\alpha = 1/(2{\cdot}0.02^2) = 1250$. Now, suppose that in reality H_0 is indeed true (i.e., the true distribution is $P_{\theta_I} = \mathrm{Bernoulli}(\theta_I)$ where $\theta_I = 0.5$), and the data are perturbed in such a way that $x_1, \ldots, x_n$ behave like i.i.d. samples from $\mathrm{Bernoulli}(\theta^o)$ where $\theta^o = 0.51$ (i.e., the observed data distribution is $P_o = \mathrm{Bernoulli}(\theta^o)$). Figure 2 (top left) shows the probability of

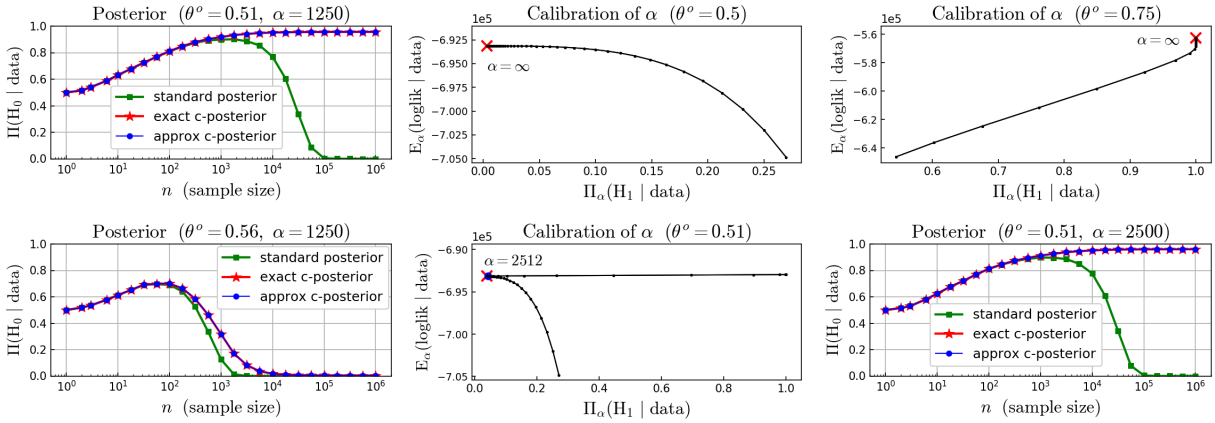


Figure 2: Bernoulli example. Left: Results using *a priori* choice of $\alpha = 1250$, averaged over 1000 datasets, for $\theta^o = 0.51$ and $\theta^o = 0.56$. Center and right: α calibration curves for $\theta^o \in \{0.5, 0.51, 0.75\}$, and results using the data-driven choice of $\alpha = 2500$ when $\theta^o = 0.51$.

H_0 under the standard posterior, the exact c-posterior, and the approximate c-posterior (Equations 3.3, 3.1, and 3.2, respectively), for increasing values of the sample size n.

When n is small, there is not enough power to distinguish between 0.5 and 0.51, so the standard posterior favors H_0 at first (due to the Bartlett–Lindley effect), but as n increases, eventually the posterior probability of H_0 goes to 0. (So, when n is large, the standard posterior is not robust to this perturbation.) The c-posterior behaves the same way as the standard posterior when n is small, but as n increases, the c-posterior probability of H_0 remains high, as desired—thus, the c-posterior remains robust for large n. The approximate c-posterior is so close to the exact c-posterior that the plots are visually indistinguishable.

What if the departure from H_0 is significantly larger than our chosen tolerance of $\varepsilon = 0.02$? Does the c-posterior more strongly favor H_1 in such cases, as it should? Indeed, it does. Figure 2 (bottom left) shows the results when $\theta^o = 0.56$. In this case, the c-posterior behaves more like the standard posterior, favoring H_1 when n is sufficiently large.

4 Calibration curve technique for choosing α

If we have no *a priori* basis for choosing α, then the following graphical criterion can help to make a data-driven choice. Let $f(\alpha)$ be a measure of fit to the data and let $g(\alpha)$ be a measure of effective complexity—specifically, we use the posterior expected log likelihood for $f(\alpha)$, and posterior expected model complexity for $g(\alpha)$. As α ranges from 0 to ∞, $(g(\alpha), f(\alpha))$ traces out a curve in $\mathbb{R}^2$, and the technique is to choose a point on this curve that achieves a good fit with low complexity.

To illustrate on the toy Bernoulli example, we define $f(\alpha) = \int \big(\log p(x_{1:n}|\theta)\big)\Pi_\alpha(d\theta|x_{1:n})$ to quantify fit to the data and $g(\alpha) = \Pi_\alpha(\mathrm{H}_1|x_{1:n})$ to quantify effective complexity, where $\Pi_\alpha(d\theta|x_{1:n}) \propto p(x_{1:n}|\theta)^{\zeta_n}\Pi(d\theta)$ is the power posterior; see Supplement S7.1 for formulas. Figure 2 shows the resulting calibration curves for three datasets of size $n = 10^6$, generated (i) when H_0 is true and there is no perturbation ($\theta^o = 0.5$), (ii) when H_0 is true and there is a small perturbation ($\theta^o = 0.51$), and (iii) when H_1 is true and distance from 0.5 is large ($\theta^o = 0.75$). In each case, the curve goes from lower fit to higher fit as α increases. The distinction between "small" and "large" distance depends on the choice of prior—e.g., $\theta^o = 0.51$ is close to 0.5 relative to typical samples from our prior of $\theta|\mathrm{H}_1 \sim \mathrm{Uniform}(0, 1)$.

The three calibration curves in Figure 2 illustrate common patterns. Case (i): When there is no perturbation from H_0, the best fit is obtained with very low complexity at the terminus $\alpha = \infty$. This suggests choosing $\alpha = \infty$, which would make the c-posterior concentrate at the true value in this case. Case (ii): When there is a small perturbation from H_0 ($\theta^o = 0.51$), the fit increases dramatically at first while maintaining low complexity, then the curve reaches a cusp at $\alpha \approx 2500$ and levels off, with more modest increases in fit at the cost of greater complexity. This suggests choosing $\alpha \approx 2500$. The curve sits near the cusp for a large range of α values from around 1200 to 4000, e.g., the blue dot indicates $\alpha = 1250$, our *a priori* choice. Any value of α in this range yields similar results (e.g., see Figure 2 bottom right compared to top left). Case (iii): When the distance from H_0 is very large ($\theta^o = 0.75$), there is no cusp in the curve, and a good fit can only be obtained

at higher complexity. This curve suggests choosing $\alpha = \infty$, in which case the c-posterior would concentrate at H_1. This makes sense since the distance from H_0 is so large that explaining it by a perturbation is implausible. Thus, the calibration curve can help decide how much coarsening is needed, if any.

5 Mixture models and clustering

Consider a finite mixture model, $X_1, \ldots, X_n | w, \varphi$ i.i.d. $\sim \sum_{i=1}^{K} w_i f_{\varphi_i}(x)$ with mixture weights w, component parameters φ, and family of component distributions $(f_\phi : \phi \in \Phi)$. For the prior, suppose $w \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_K)$ and $\varphi_1, \ldots, \varphi_K$ i.i.d. $\sim H$. When $\gamma_i = c/K$, this model approximates a Dirichlet process mixture as $K \to \infty$ (Ishwaran and Zarepour, 2002). Mixture models of this form are widely used for clustering.

However, this type of model is not robust to misspecification of the family of component distributions. This has negative consequences in practice, since one might reasonably expect the observed data $x_1, \ldots, x_n$ to come from a finite mixture, but it is usually unreasonable to expect the component distributions to have a known parametric form. We illustrate how coarsening enables one to perform inference in a way that is robust to misspecification of the form of the component distributions.

We approximate the relative entropy c-posterior using the power posterior, defined as $\pi_\alpha(w, \varphi \mid x_{1:n}) \propto \pi(w, \varphi) \prod_{j=1}^{n} \big(\sum_{i=1}^{K} w_i f_{\varphi_i}(x_j) \big)^{\zeta_n}$ where $\zeta_n = \alpha/(\alpha + n)$. The standard MCMC algorithms for mixture models use data augmentation with latent variables $z_1, \ldots, z_n \in \{1, \ldots, K\}$ indicating which component each datapoint comes from, but the power likelihood rules out direct application of these algorithms. Antoniano-Villalobos and Walker (2013) developed an auxiliary variable algorithm for mixture model power posteriors, or reversible jump MCMC could be used (Green, 1995).

Here, we explore two algorithms: (a) a conditional coarsening algorithm and (b) an importance sampling algorithm for the power posterior. The conditional coarsening algorithm scales well, is easy to implement, and yields results similar to (but not exactly the

same as) the power posterior. It is identical to the standard data augmentation algorithm for mixtures, except that the updates to w and φ use a power likelihood.

Algorithm 5.1 (Conditional coarsening for mixture models)**.**

- *Input:* $x_1, \ldots, x_n$. *Output: Samples of* w, φ, *and component assignments* $z_1, \ldots, z_n$.
- *Initialize* $w \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_K)$ *and* $\varphi_1, \ldots, \varphi_K$ *i.i.d.* $\sim H$.
- *For iteration* $t = 1, \ldots, T$*:*

  1. *For* $j = 1, \ldots, n$*: sample* $z_j \sim \text{Categorical}(\tilde{w})$ *where* $\tilde{w}_i \propto w_i f_{\varphi_i}(x_j)$.
  2. *Sample* $w \sim \text{Dirichlet}(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_K)$ *where* $\tilde{\gamma}_i = \gamma_i + \zeta_n \sum_{j=1}^n \mathbb{1}(z_j = i)$.
  3. *For* $i = 1, \ldots, K$*: sample* $\varphi_i \sim q$ *where* $q(\varphi_i) \propto \pi(\varphi_i) \prod_{j:z_j=i} f_{\varphi_i}(x_j)^{\zeta_n}$*, or make some other update to* φ_i *that leaves* q *invariant.*

See Supplement S7.2 for the motivation behind the algorithm. In some cases, Algorithm 5.1 has difficulty escaping from local optima in which one cluster needs to be split into two or more clusters. Therefore, we add the following step between steps 1 and 2, to escape from these local optima during an initialization period that is discarded along with the burn-in. Let S and T_{init} be positive integers. Define $N_i(z) = \sum_{j=1}^n \mathbb{1}(z_j = i)$ and $k(z) = \sum_{i=1}^K \mathbb{1}(N_i(z) > 0)$.

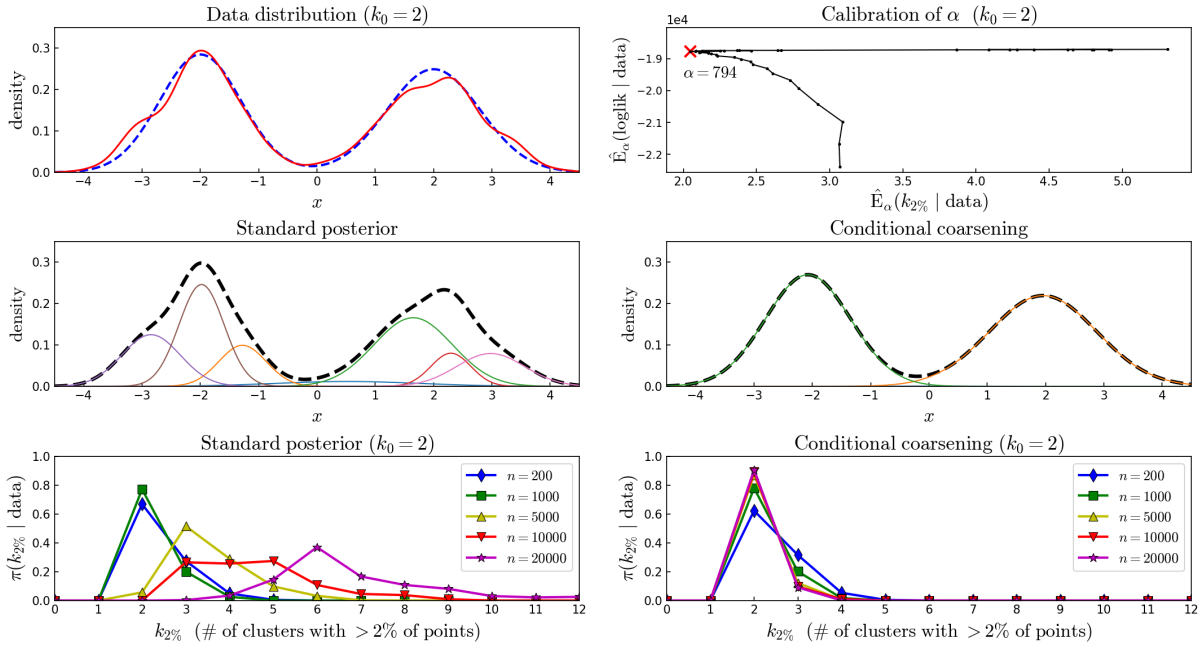
1.5. (Periodic random splits) If $t < T_{init}$ *and* t *is a multiple of* S*, then randomly split each of the* $K - k(z)$ *largest clusters into two clusters.*

More precisely, let σ such that $N_{\sigma_1}(z) \geq \cdots \geq N_{\sigma_K}(z)$, and let $K' = k(z)$. Then, for $i = 1, \ldots, \min\{K', K - K'\}$: for each j such that $z_j = \sigma_i$, update $z_j \sim \text{Uniform}\{\sigma_i, \sigma_{i+K'}\}$.

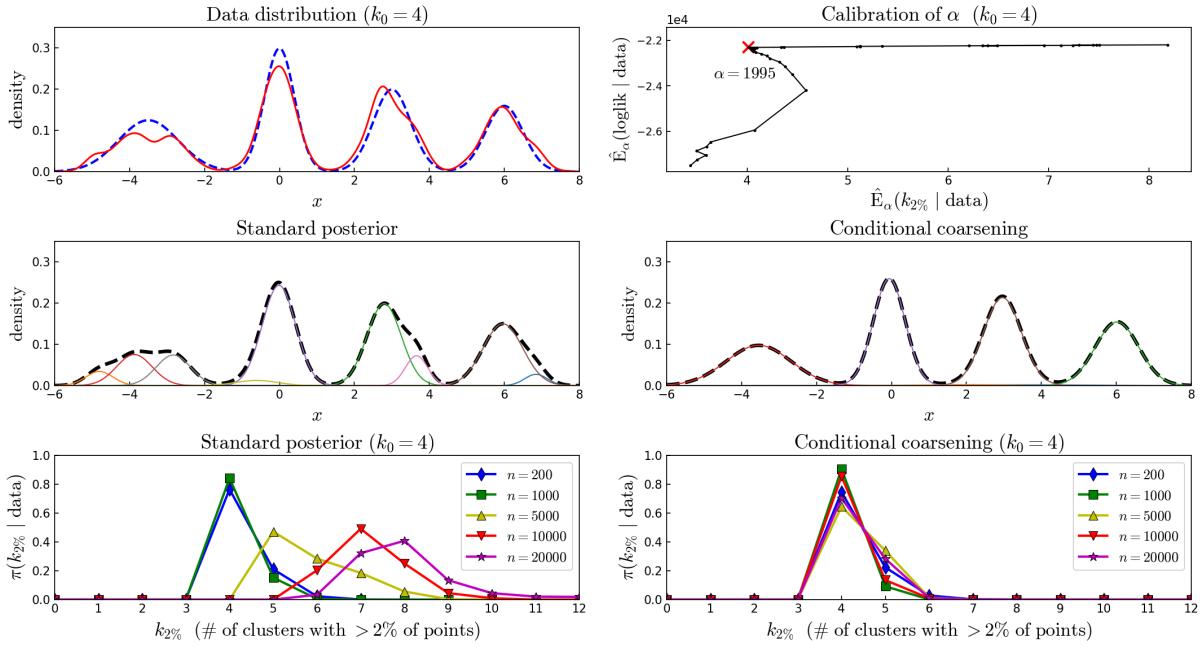
To evaluate how closely the conditional coarsening algorithm approximates the power posterior, we also consider an importance sampling (IS) algorithm; see Supplement S7.2.

5.1 Simulation example: Perturbed mixture of Gaussians

To demonstrate robustness to the form of the component distributions, we apply a univariate Gaussian mixture model to data from a perturbed Gaussian mixture. We generate the



(a) Demonstration on a perturbed mixture of $k_0 = 2$ Gaussians.



(b) Demonstration on a perturbed mixture of $k_0 = 4$ Gaussians.

Figure 3: Top left: True density (dotted blue line) and perturbed density (red line). Top right: Calibration curve for α. Middle: Mixture density (dotted black line) and components (solid colors) for typical samples from the posterior. Bottom left: The standard posterior has too many clusters as n increases. Bottom right: Coarsening yields a more accurate number of clusters.

observed data by starting with a true (idealized) distribution $P_{\theta_I} = \sum_{i=1}^{k_0} w_{0i}\mathcal{N}(\mu_{0i}, \sigma_{0i}^2)$, simulating a perturbation P_o by taking a random draw of a Dirichlet process mixture with base distribution P_{θ_I}, concentration parameter 500, and $\mathcal{N}(0, 0.25^2)$ components, and then sampling $x_1, \ldots, x_n$ i.i.d. $\sim P_o$. We illustrate with two examples: (a) a two-component mixture with $\mu_0 = (-2, 2)$, $\sigma_0 = (.7, .8)$, and $w_0 = (.5, .5)$, and (b) a four-component mixture with $\mu_0 = (-3.5, 0, 3, 6)$, $\sigma_0 = (.8, .4, .5, .5)$, and $w_0 = (.25, .3, .25, .2)$; see Figure 3.

For the model parameters, we use $K = 20$, $\gamma_1 = \cdots = \gamma_K = 0.5/K$, and define the prior H on the component means and variances as $\mu_i \sim \mathcal{N}(m, \ell^{-1})$ and $\sigma_i^2 \sim \text{InverseGamma}(a, b)$ independently with $m = 0$, $\ell = 1/5^2$, $a = 1$, and $b = 1$, where the component densities are $f_{\mu_i,\sigma_i^2}(x) = \mathcal{N}(x|\mu_i, \sigma_i^2)$. To implement Algorithm 5.1, we define $\varphi_i = (\mu_i, \sigma_i^2)$ and for step 3 of the algorithm, we use power-likelihood Gibbs updates to μ_i and σ_i^2, specifically:

3. For $i = 1, \ldots, K$, sample

- $\mu_i \sim \mathcal{N}(\tilde{m}, \tilde{\ell}^{-1})$ *where* $\tilde{\ell} = \ell + \zeta_n N_i(z)/\sigma_i^2$, $\tilde{m} = (m\ell + \zeta_n \sum_{j:z_j=i} x_j/\sigma_i^2)/\tilde{\ell}$, *and*
- $\sigma_i^2 \sim \text{InverseGamma}(\tilde{a}, \tilde{b})$ *where* $\tilde{a} = a + \frac{1}{2}\zeta_n N_i(z)$ *and* $\tilde{b} = b + \frac{1}{2}\zeta_n \sum_{j:z_j=i}(x_j - \mu_i)^2$.

Recall that $N_i(z) = \sum_{j=1}^n \mathbb{1}(z_j = i)$. In each run of Algorithm 5.1, we use $T = 10^4$ iterations with a burn-in period of $T_{\text{burn}} = 1000$. Periodic random splits (step 1.5) are performed using $S = 10$ and $T_{\text{init}} = 500$. Samples from the standard posterior are obtained by setting ζ_n to 1. For coarsening, we use $\zeta_n = \alpha/(\alpha + n)$ with α chosen as follows.

In this type of model, posterior samples often have one or more tiny "extra" clusters. To focus on the larger clusters, we use the statistic $k_{2\%}(z) = \sum_{i=1}^K \mathbb{1}(N_i(z) > 0.02n)$ (i.e., the number of clusters with more than 2% of the points) to quantify the number of nonnegligible clusters, for a given assignment vector z. To choose α, we plot the calibration curve with $f(\alpha) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \log p(x_{1:n}|w^{(t)}, \varphi^{(t)})$ to assess fit (where $p(x_j|w, \varphi) = \sum_{i=1}^K w_i f_{\varphi_i}(x_j)$) and $g(\alpha) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} k_{2\%}(z^{(t)})$ to assess effective complexity, where $(w^{(t)}, \varphi^{(t)}, z^{(t)})$ for $t = 1, \ldots, T$ are the posterior samples obtained using Algorithm 5.1, and $\mathcal{T} = \{T_{\text{burn}}+1, \ldots, T\}$.

Figure 3(a,b) shows the calibration curves for the $k_0 = 2$ and $k_0 = 4$ examples, with $n = 10^4$ data points. In both examples, there is a clear cusp at a point of good fit and low

complexity. In the $k_0 = 2$ example, the curve is near the cusp when α is around 800 to 1000, and the tip is at $\alpha \approx 800$; thus, we choose $\alpha = 800$ in this example. In the $k_0 = 4$ example, a wide range of α values from 800 to 2000 are near the cusp, with the tip at $\alpha \approx 2000$; thus, we pick $\alpha = 2000$ in this case.

To assess performance, for both the two- and four-component examples, for each $n \in \{200, 1000, 5000, 10000, 20000\}$, we generated five independent datasets of size n. On each dataset, for the standard posterior and for conditional coarsening, Algorithm 5.1 was run using the settings above. The IS algorithm was also run using the same settings, and yielded results similar to conditional coarsening; see Supplement S7.2.

In each of Figure 3(a) and (b), the middle row shows the mixture density $\sum_{i=1}^{K} w_i f_{\mu_i,\sigma_i^2}(x)$ and the individual weighted components $w_i f_{\mu_i,\sigma_i^2}(x)$ for typical posterior samples when $n = 20000$. Samples from the standard posterior more closely fit the perturbed distribution P_o, and they have several more nonnegligible components than the true mixture P_{θ_I}. Meanwhile, typical samples using the coarsened approach more closely match the true mixture P_{θ_I} in terms of the number of nonnegligible components, as well as the weights, locations, and scales of the components.

The bottom row in each of Figure 3(a) and (b) shows the posterior on $k_{2\%}$ (the number of clusters containing more than 2% of the points), averaged over the five datasets. The standard posterior tends to introduce more clusters as n increases, in order to fit the observed data distribution P_o. Meanwhile, the coarsened approach shows strong support for the true number of nonnegligible clusters, no matter how large n becomes.

In summary, when there is a perturbation, the coarsened approach yields more accurate inferences for the true (unperturbed) mixture parameters.

5.2 Application: Robust clustering for flow cytometry

Flow cytometry is a high-throughput technology for measuring the properties of individual cells in a sample of biological material. Typically, in each sample, tens of thousands of

individual cells are measured with respect to around 3 to 20 properties. In flow cytometry data, cells from distinct populations tend to fall into clusters; see Figure 4. Discovery and characterization of cell populations by clustering is one of the primary tasks performed with this type of data. Traditionally, this clustering is performed manually by defining piecewise linear boundaries between regions; this is known as "gating". Since manual gating is labor intensive and subjective, several automated clustering algorithms have been developed, and the Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP) challenges were established to evaluate the performance of these methods on benchmark datasets with ground truth determined by manual gating (Aghaeepour et al., 2013).

We consider 12 of these benchmark datasets, originally from a longitudinal study of graft-versus-host disease (GvHD) in patients undergoing blood or marrow transplantation (Brinkman et al., 2007). Each dataset corresponds to one blood draw from one patient. The objective of the study was to understand how various cell populations differed between patients who developed GvHD and patients who did not. Separating distinct populations of cells is the first step in the analysis of this data.

The difficulty is that the populations are not well-approximated by any parametric distribution, and further, the number of populations is not known in advance. Consequently, using a model such as a mixture of Gaussians yields poor results, since many Gaussians are needed to fit each population; see Figure 4 (row 2). Some previous algorithms for flow cytometry have dealt with this problem by performing a *post hoc* step in which multiple clusters are grouped together (Finak et al., 2009; Aghaeepour et al., 2011). Ideally, one would use a nonparametric model for the component distributions, but this is computationally intensive due to the large number of multivariate data points in each sample.

We explore a coarsening approach to robust clustering for flow cytometry data, using a multivariate Gaussian mixture model. For the model parameters, in the same notation as at the beginning of Section 5, we use $K = 20$, $\gamma_1 = \cdots = \gamma_K = 0.5/K$, and component parameter priors $\mu_i \sim \mathcal{N}(m, L^{-1})$ and $\Lambda_i \sim \text{Wishart}(V, \nu)$ independently, where the component densities are $f_{\mu_i,\Lambda_i}(x) = \mathcal{N}(x|\mu_i, \Lambda_i^{-1})$ for $x \in \mathbb{R}^d$. We set the hyperparameters in a data-

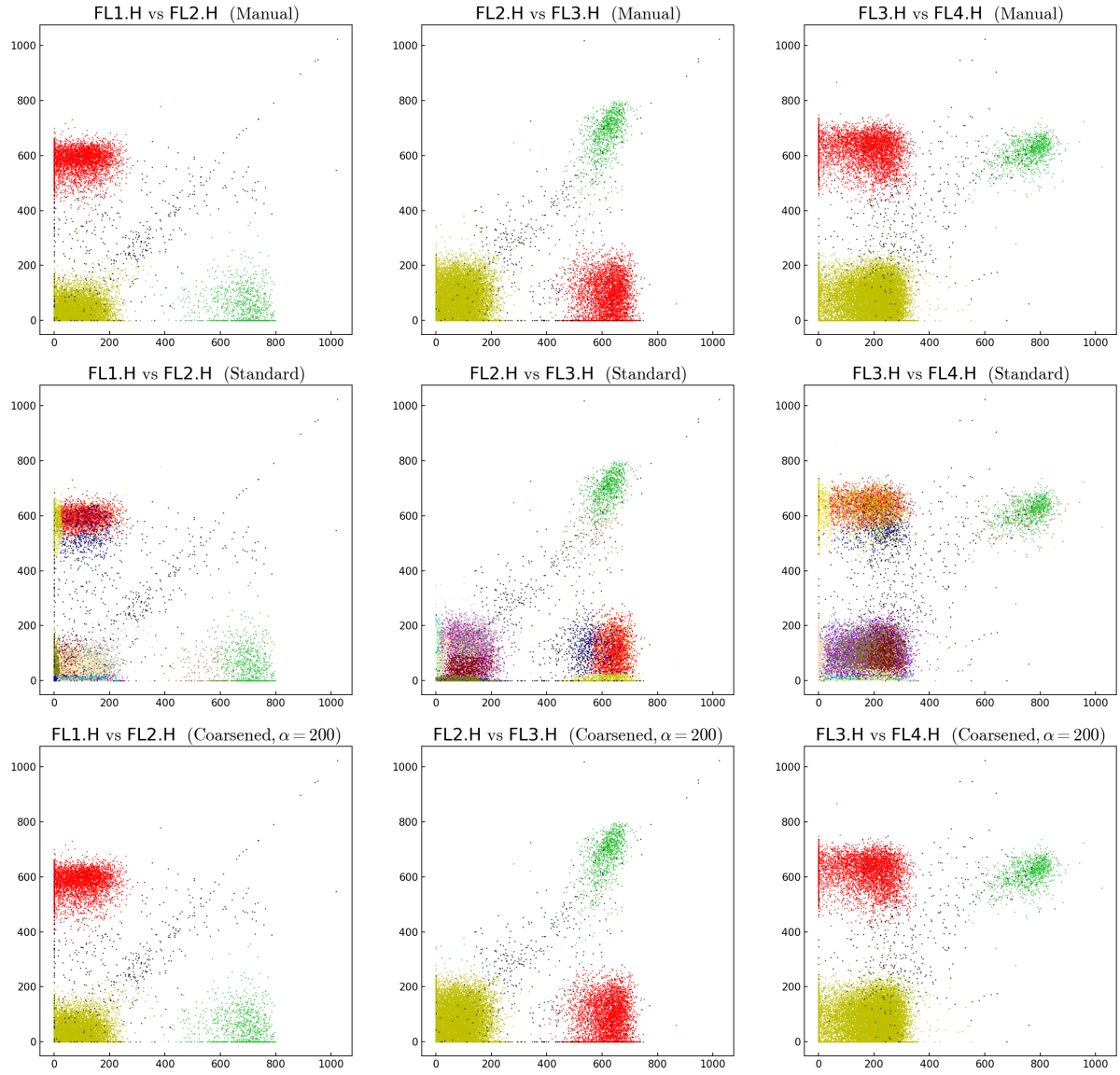


Figure 4: Flow cytometry clustering results on FlowCAP-I GvHD dataset #10 ($n = 23377$, $d = 4$). The four dimensions are FL1.H, FL2.H, FL3.H, and FL4.H, which measure selected antibodies; three two-dimensional projections are shown. Row 1: Expert manual gating identifies three populations (clusters) of cells. Each point is one cell, and the colors indicate cluster labels, with black indicating cells not labeled by the expert. Row 2: The standard posterior yields far too many clusters — on this dataset, posterior samples typically have 13 clusters that contain more than 2% of the points, each. Row 3: Conditional coarsening very closely matches the manual ground truth (average F-measure = 0.998 in this case). In rows 2–3, the clusters shown are the z^* assignments from the last iteration of the algorithm.

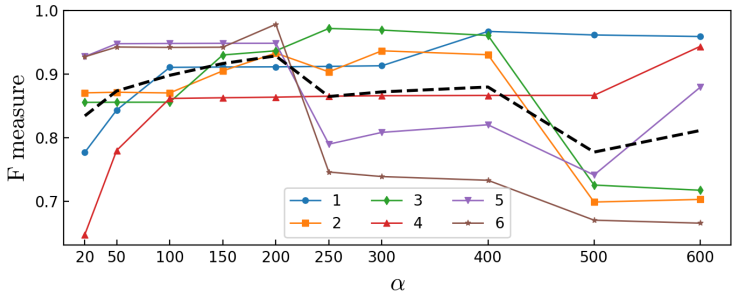


Figure 5: Calibration of α on the training set (GvHD datasets 1–6). The average F-measure is shown for each α and each dataset. The black dotted line is the overall average for each α.

dependent way: given input data $x_1, \ldots, x_n \in \mathbb{R}^d$, we choose prior mean $m = \frac{1}{n}\sum_{j=1}^n x_j$, prior precision matrix $L = \left(\frac{1}{n}\sum_{j=1}^n (x_j - m)(x_j - m)^\intercal\right)^{-1}$, degrees of freedom $\nu = d$, and scale matrix $V = L/\nu$. Algorithm 5.1 is implemented by defining $\varphi_i = (\mu_i, \Lambda_i)$ and using power-likelihood Gibbs updates to μ_i and Λ_i for step 3 of the algorithm:

3. For $i = 1, \ldots, K$, sample

- $\mu_i \sim \mathcal{N}(\tilde{m}, \tilde{L}^{-1})$ *where* $\tilde{L} = L + \zeta_n N_i(z)\Lambda_i$, $\tilde{m} = \tilde{L}^{-1}(Lm + \zeta_n \Lambda_i \sum_{j:z_j=i} x_j)$, *and*
- $\Lambda_i \sim \text{Wishart}(\tilde{V}, \tilde{\nu})$ *where* $\tilde{\nu} = \nu + \zeta_n N_i(z)$, $\tilde{V}^{-1} = V^{-1} + \zeta_n \sum_{j:z_j=i}(x_j - \mu_i)(x_j - \mu_i)^\intercal$.

In each iteration of the algorithm, we compute $z_j^* = \text{argmax}_i\, w_i f_{\mu_i,\Lambda_i}(x_j)$ for $j = 1, \ldots, n$, the most likely component assignments based on the parameter values at that iteration.

In each run of Algorithm 5.1, we use $T = 4000$, $T_{\text{burn}} = 2000$, $S = 10$, and $T_{\text{init}} = 400$. Setting ζ_n to 1 yields the standard posterior, and for coarsening we use $\zeta_n = \alpha/(\alpha + n)$. To choose α, we split the data into a training set (datasets 1–6) and a test set (datasets 7–12). The performance metric used in FlowCAP-I is F-measure, a similarity score between any two partitions $\mathcal{A}$ and $\mathcal{B}$ of $\{1, \ldots, N\}$, defined as

$$F(\mathcal{A}, \mathcal{B}) = \sum_{A \in \mathcal{A}} \frac{|A|}{N} \max_{B \in \mathcal{B}} \frac{2|A \cap B|}{|A| + |B|}.$$

For a range of α values, for each training dataset, we run Algorithm 5.1 and at each iteration we compute $F(\mathcal{A}, \mathcal{B})$ with $\mathcal{A}$ as the manual ground truth and $\mathcal{B}$ as the partition induced by

Table 1: Average F-measures on the flow cytometry test set (GvHD datasets 7–12).

	7	8	9	10	11	12
Standard	0.532	0.478	0.619	0.453	0.542	0.585
Coarsened	0.667	0.875	0.931	0.998	0.989	0.993

z^*. In each dataset, a small fraction of cells were not labeled by the human expert; these unlabeled cells are included when running the algorithm, and excluded when computing the F-measure. Figure 5 shows the average F-measure for each of these runs, excluding burn-in. Averaging over the six training datasets, the best performance is obtained at $\alpha = 200$; thus, we set $\alpha = 200$ to evaluate performance on the test datasets.

Table 1 shows the average F-measures on the test set (datasets 7–12), using the same algorithm settings as above, comparing z^* against ground truth as before. The standard posterior performs very poorly, whereas the coarsening results are comparable to the best performance obtained by algorithms tailored to flow cytometry clustering (Aghaeepour et al., 2013). Of datasets 7–12, coarsening has the most difficulty on 7, but interestingly, if we increase α to 500, then the F-measure increases to 0.937 and the resulting cluster assignments closely resemble the ground truth; see Figure 6. This suggests that even better performance may be possible with an improved method of choosing α for each dataset.

6 Autoregressive models of unknown order

In this section, we apply the c-posterior to perform inference for the order of an autoregressive model in a way that is robust to misspecification of the structure of the model, such as time-varying noise. This demonstrates how the robustified marginal likelihood can be computed in closed form when using conjugate priors, and provides some insight into why coarsening works. Consider an AR(k) model, that is, a kth-order autoregressive model: $X_t = \sum_{\ell=1}^{k} \theta_\ell X_{t-\ell} + \varepsilon_t$ for $t = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ and $X_t = 0$ for $t \leq 0$. Let $\pi(k)$ be a prior on the order k, let $\theta_1, \ldots, \theta_k | k$ i.i.d. $\sim \mathcal{N}(0, \sigma_0^2)$, and for

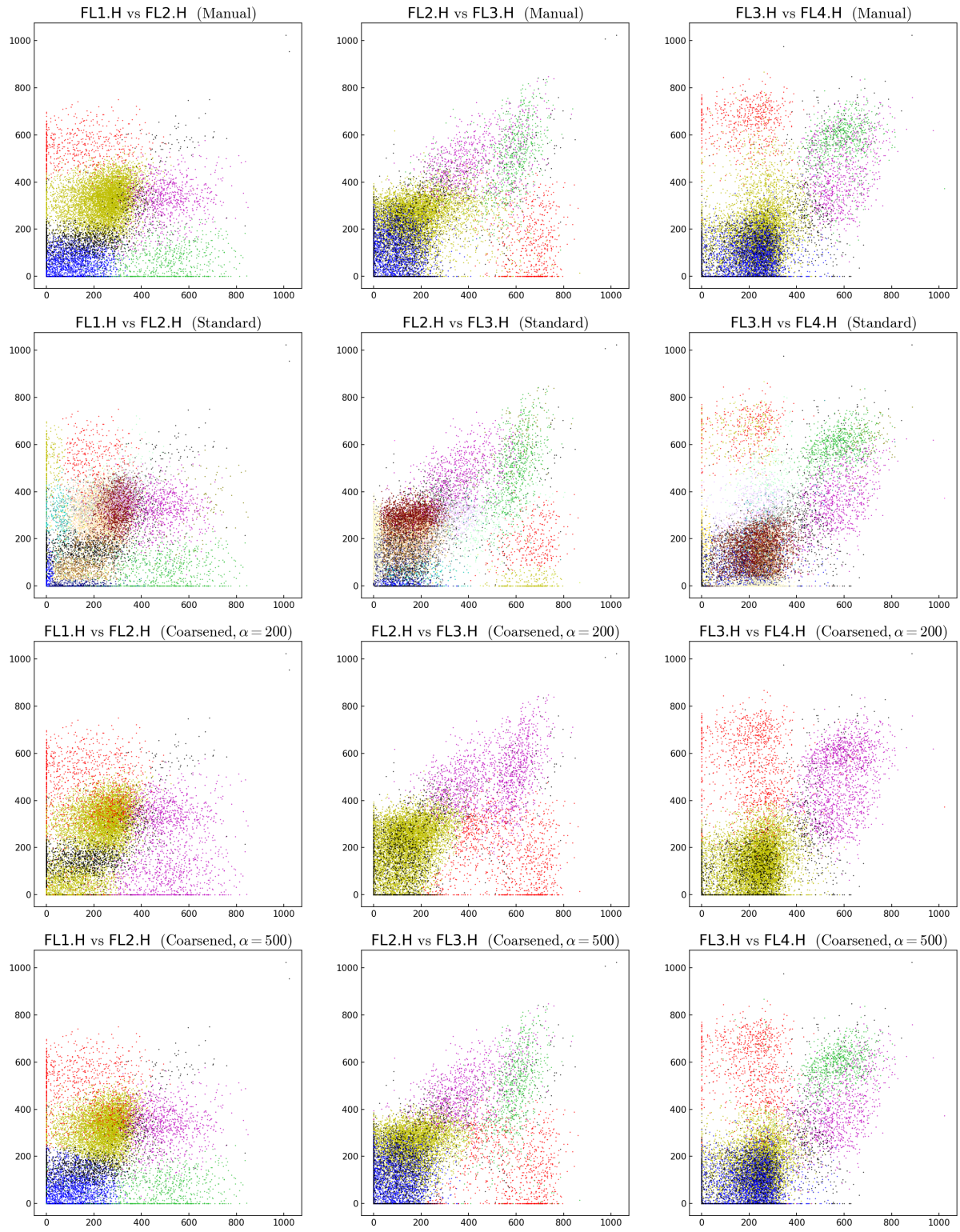


Figure 6: Flow cytometry clustering results on FlowCAP-I GvHD dataset #7 ($n = 13773$, $d = 4$). Row 1: Ground truth clusters from expert labeling. Row 2: Standard posterior. Rows 3-4: Conditional coarsening with $\alpha \in \{200, 500\}$. See the text and Figure 4 for more information.

simplicity, assume σ^2 is known.

To obtain robustness to perturbations that are small with respect to relative entropy rate, we employ a c-posterior for time-series (see Supplement S6.1 for details). Since $\theta|k$ has been given a conjugate prior, we can analytically compute the resulting marginal power likelihood as described in Section 2.1.1 with power $\zeta_n = \alpha/(\alpha+n)$,

$$\begin{aligned} L_\alpha(k; x_{1:n}) &:= \int_{\mathbb{R}^k} p(x_{1:n}|\theta, k)^{\zeta_n} \pi(\theta|k) d\theta \\ &= \int_{\mathbb{R}^k} \Big(\prod_{t=1}^{n} \mathcal{N}\big(x_t \,\big|\, \textstyle\sum_{\ell=1}^{k} \theta_\ell x_{t-\ell},\, \sigma^2\big) \Big)^{\zeta_n} \mathcal{N}(\theta \mid 0, \sigma_0^2 I_{k\times k}) d\theta \\ &= \frac{\exp(\frac{1}{2}\zeta_n^2 v^{\mathsf{T}} \Lambda^{-1} v)}{\sigma_0^k |\Lambda|^{1/2}} \mathcal{N}(x_{1:n} \mid 0, \sigma^2 I_{n\times n})^{\zeta_n} \end{aligned}$$

where $\Lambda = \zeta_n M + \sigma_0^{-2} I_{k\times k}$, $M_{ij} = \sum_{t=1}^{n} x_{t-i} x_{t-j}/\sigma^2$, $v_i = \sum_{t=1}^{n} x_t x_{t-i}/\sigma^2$, and $x_t = 0$ for $t \le 0$. This, in turn, can be used to compute a robustified posterior on the model order k, defined as $\pi_\alpha(k|x_{1:n}) \propto L_\alpha(k; x_{1:n})\pi(k)$.

To demonstrate empirically, we generate data from a process that is close to AR(4) but exhibits time-varying noise that cannot be captured by the model:

$$x_t = \sum_{\ell=1}^{4} \theta_\ell x_{t-\ell} + \varepsilon_t + \tfrac{1}{2}\sin t \tag{6.1}$$

where $\theta = (1/4, 1/4, -1/4, 1/4)$, ε_t i.i.d. $\sim \mathcal{N}(0,1)$, and $x_t = 0$ for $t \le 0$. We apply the model above to such data, and compare the standard Bayesian approach to the coarsened approach. For the model parameters, we set $\sigma^2 = 1$ to match the true value, we set $\sigma_0^2 = 1$, and we use a Geometric(0.1) prior on k (i.e., $\pi(k) = 0.9^k 0.1$ for $k \in \{0, 1, 2, \ldots\}$).

To choose α, we use the calibration technique described in Section 4. Specifically, for a range of α values, we compute $f(\alpha) = \sum_k \big(\log p(x_{1:n}|k)\big)\pi_\alpha(k|x_{1:n})$ as a measure of fit to the data (noting that $\log p(x_{1:n}|k) = \log L_\infty(k; x_{1:n})$), and $g(\alpha) = \sum_k k\, \pi_\alpha(k|x_{1:n})$ as a measure of effective complexity. Figure 7 (top right) shows the resulting calibration curve, for a dataset of size $n = 10^4$. The fit increases sharply until a cusp is reached at $\alpha \approx 250$, whereupon the curve levels off; in fact, a wide range of α values from around 200 to 600 are very near the cusp. This suggests choosing $\alpha = 250$.

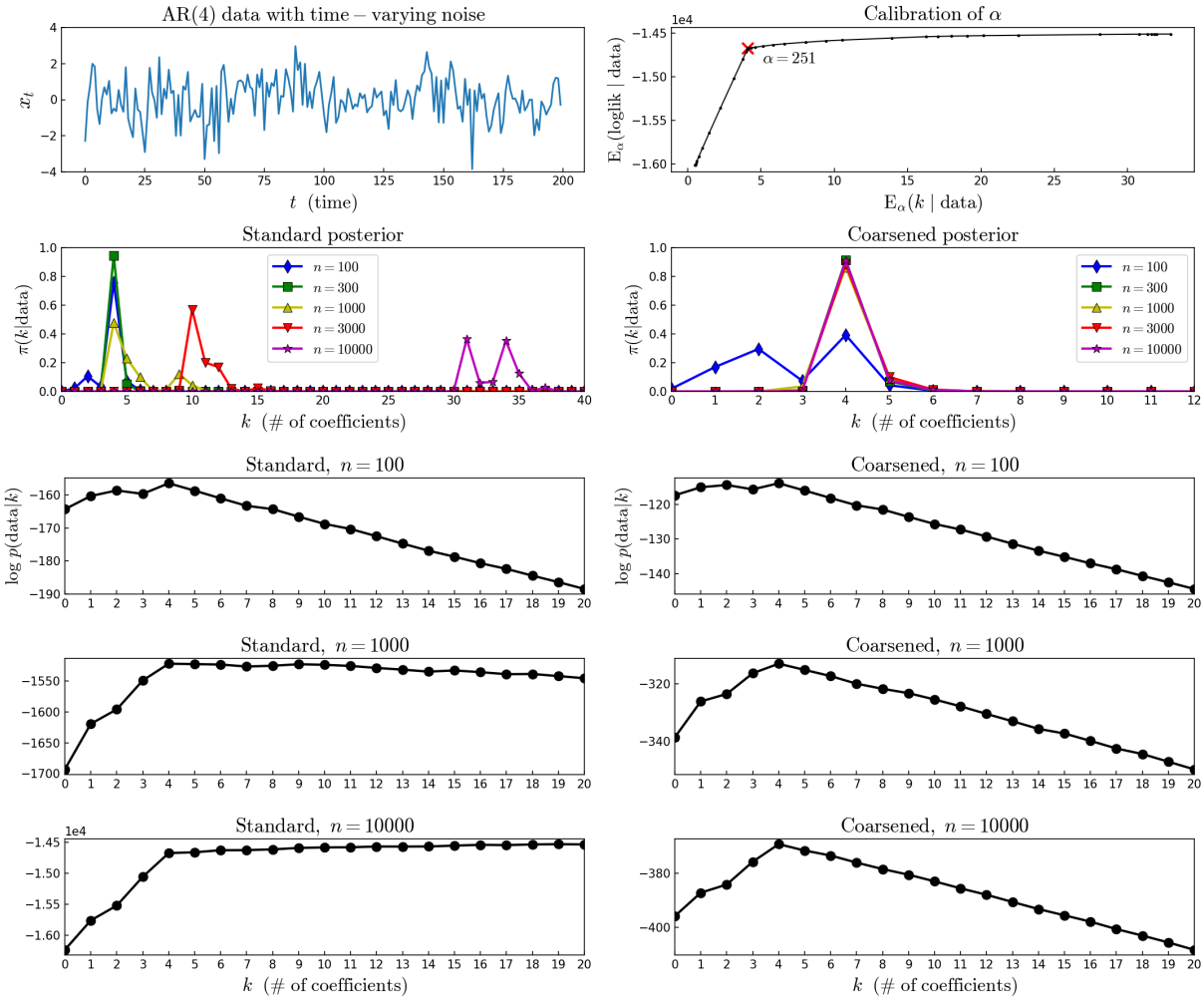


Figure 7: Autoregression example. Row 1 (left): Data from the perturbed AR(4) process in Equation 6.1. Row 1 (right): α calibration curve, when $n = 10^4$. Row 2: Posterior distributions on k. Note that the standard posterior significantly overestimates k as n grows, whereas the c-posterior strongly favors the true value of k. Rows 3-5: Log marginal likelihood of AR(k) model (standard and coarsened) for $k = 0, 1, \ldots, 20$, on increasing amounts of data from this process.

Figure 7 (rows 2–5) compares the standard posterior to the c-posterior with $\alpha = 250$, as n increases. Due to the misspecification, the standard posterior puts its mass on values of k much greater than the true value of 4, when n gets sufficiently large. Meanwhile, the c-posterior stabilizes to a distribution on k favoring $k = 4$. When $n < \alpha$, the standard and coarsened approaches yield similar results, but as n grows larger they differ markedly.

This pattern is typical of the log marginal likelihood when comparing models of increasing complexity. From the Laplace approximation, we see that more complex models are penalized via a term of the form $-\frac{1}{2}t_k \log n$ where t_k is the dimension of the parameter for model k (see Supplement S4), e.g., $t_k = k$ for the AR(k) model above. This penalty is visible in the linear decline exhibited in the $n = 100$ plot. As n increases, this complexity penalty increases proportionally to only $\log n$, and thus it becomes overwhelmed by the main term of order n involving the log likelihood at the maximum likelihood estimator within model k. When n is sufficiently large, the following pattern emerges, as seen in the $n = 10000$ plot for the standard approach: for model complexity values k that are too small, there is a clear lack of fit, and as k increases the log marginal likelihood increases rapidly until the model can fairly closely approximate the data distribution, at which point it plateaus, increasing only slightly after that as only fine grain improvements can be made.

From this perspective, the reason why the coarsened marginal likelihood "works" is that when n is large, it maintains a balance between the model complexity penalty and the main log-likelihood term.

7 Variable selection in linear regression

Consider the following spike-and-slab model for variable selection:

$$
\begin{aligned}
&W \sim \text{Beta}(r, s) \\
&\beta_j \sim \mathcal{N}(0, 1/L_0) \text{ with probability } W, \text{ otherwise } \beta_j = 0, \text{ for each } j = 1, \ldots, p \\
&\lambda \sim \text{Gamma}(a, b) \\
&Y_i | \beta, \lambda \sim \mathcal{N}(\beta^{\mathsf{T}} x_i, 1/\lambda) \text{ independently for } i = 1, \ldots, n.
\end{aligned}
\tag{7.1}
$$

Models of this type are often used to infer which covariates $x_{i1}, \ldots, x_{ip}$ are predictive of the target variable y_i, by considering which coefficients β_j have a high posterior probability of being nonzero. This provides valuable insight into the relationships present in the data generating process. However, usually, it is unlikely that the data exactly follow

the $\mathcal{N}(\beta^{\mathsf{T}} x_i, 1/\lambda)$ form, and although the model exhibits some robustness to departures from normality, it is not robust to departures from the assumed form of the mean function $\beta^{\mathsf{T}} x_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. We demonstrate how the c-posterior provides robustness to misspecification of this type. This example also shows how Gibbs sampling can be used with power posteriors when conditionally conjugate priors have been chosen.

In a regression setting, it is natural to use the c-posterior based on conditional relative entropy. Just as before, this can be approximated by the power posterior obtained by raising the likelihood to $\zeta_n = \alpha/(\alpha + n)$ (see Supplement S6.2). If we first integrate W out of the model, the resulting power posterior is $\pi_\alpha(\beta, \lambda | y_{1:n}) \propto p(y_{1:n} | \beta, \lambda)^{\zeta_n} \pi(\beta, \lambda)$. Due to the use of conditionally conjugate priors, the full conditionals for β_j and λ can be derived in closed form by standard calculations, and we obtain the following simple algorithm.

Algorithm 7.1 (Gibbs sampler for coarsened variable selection)**.**

- *Input:* $(x_1, y_1), \ldots, (x_n, y_n)$ *where* $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. *Output: Samples of* β *and* λ.
- *Initialize* $\beta = 0 \in \mathbb{R}^p$ *and* $\lambda = 1$.
- *For iteration* $t = 1, \ldots, T$*:*

1. *Sample* $\lambda \sim \text{Gamma}(A, B)$ *where* $A = a + \frac{1}{2} n \zeta_n$ *and* $B = b + \frac{1}{2} \zeta_n \sum_{i=1}^n (y_i - \beta^{\mathsf{T}} x_i)^2$.
2. *For* $j = 1, \ldots, p$*:*
   (a) *Sample* $z \sim \text{Bernoulli}(1 - q)$ *with* $q = \Pi_\alpha(\beta_j = 0 \mid \beta_{-j}, \lambda, y_{1:n})$ *as in Eqn 7.2.*
   (b) *If* $z = 0$*, set* $\beta_j = 0$*. If* $z = 1$*, sample* $\beta_j \sim \mathcal{N}(M, L^{-1})$ *with* M, L *as in Eqn 7.2.*

We parameterize Gamma distributions by shape and rate, i.e., $\text{Gamma}(\lambda | A, B) \propto \lambda^{A-1} e^{-B\lambda} \mathbb{1}(\lambda > 0)$. Under $\pi_\alpha(\beta, \lambda | y_{1:n})$, the conditional probability that $\beta_j = 0$ is

$$\Pi_\alpha(\beta_j = 0 \mid \beta_{-j}, \lambda, y_{1:n}) = \left(1 + \sqrt{L_0/L}\, \exp\left(\tfrac{1}{2} L M^2\right) \frac{r + \sum_{\ell \neq j} \mathbb{1}(\beta_\ell \neq 0)}{s + \sum_{\ell \neq j} \mathbb{1}(\beta_\ell = 0)}\right)^{-1} \tag{7.2}$$

where $L = L_0 + \lambda \zeta_n \sum_{i=1}^n x_{ij}^2$, $M = (\lambda \zeta_n / L) \sum_{i=1}^n \delta_i x_{ij}$, and $\delta_i = y_i - \sum_{\ell \neq j} \beta_\ell x_{i\ell}$.

It is natural to wonder whether using a robust prior would reduce or eliminate the need for coarsening. A leading example of a robust prior is the *mixture of g priors* proposed by Bayarri et al. (2012) for variable selection in linear regression. It turns out that such a prior

does not reduce the need for coarsening when there is a perturbation from the model. A robust prior provides robustness to the choice of prior, but not robustness to the choice of likelihood. This is demonstrated empirically in Section 7.1. To implement the mixture of g priors, we use the MCMC algorithm from Hoff (2009) (Sec. 9.3.2), coupled with Metropolis updates for g using $\mathcal{N}(g, (4n)^2)$ proposals. For the prior on g, in the notation of Bayarri et al. (2012) (Sec. 3.4), we use $a = 1$, $b = 1$, and $\rho_i = (k_i + 1)^{-1}$ since $k_0 = 0$. To make the prior on the number of nonzero coefficients k match the prior $\pi(k)$ implied by the model in Equation 7.1, we modify their suggested prior on the inclusion variables $z \in \{0, 1\}^p$ to be $\pi(z) = \pi(k)/\binom{p}{k}$ where $k = \sum_{j=1}^{p} z_j$.

7.1 Simulation example: Quadratic perturbation

Consider a simulation example where the true function of interest is $-1 + 4x_{i2}$, but the observed data have been slightly perturbed such that $y_i = -1 + 4x_{i2} + \frac{1}{4}x_{i2}^2 + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, 1)$. Set $x_{i1} = 1$ to accommodate the intercept, and suppose there are five covariates $x_{i2}, \ldots, x_{i6}$ distributed according to a multivariate skew-normal distribution (Azzalini and Capitanio, 1999) that has been centered and scaled so that each covariate has zero mean and unit variance: $X_{ij} = (\tilde{X}_{ij} - \mathbb{E}\tilde{X}_{ij})/\sigma(\tilde{X}_{ij})$ for $j = 2, \ldots, 6$, where $\tilde{X}_i \sim \mathcal{SN}_5(\Omega, \tilde{a})$ with shape $\tilde{a} = (0.6, 2.7, -3.3, -4.9, -2.5)$ and scale matrix

$$\Omega = \begin{pmatrix} 1.0 & -0.89 & 0.93 & -0.91 & 0.98 \\ -0.89 & 1.0 & -0.94 & 0.97 & -0.91 \\ 0.93 & -0.94 & 1.0 & -0.96 & 0.97 \\ -0.91 & 0.97 & -0.96 & 1.0 & -0.93 \\ 0.98 & -0.91 & 0.97 & -0.93 & 1.0 \end{pmatrix}.$$

The $\tilde{a}$ and Ω above were randomly generated; there is nothing particularly special about them, except that Ω was chosen so that the covariates would be fairly strongly correlated. Figure 8 (top left) shows a scatterplot of y_i versus x_{i2} for 200 samples, as well as the perturbed mean as a function of x_{i2}.

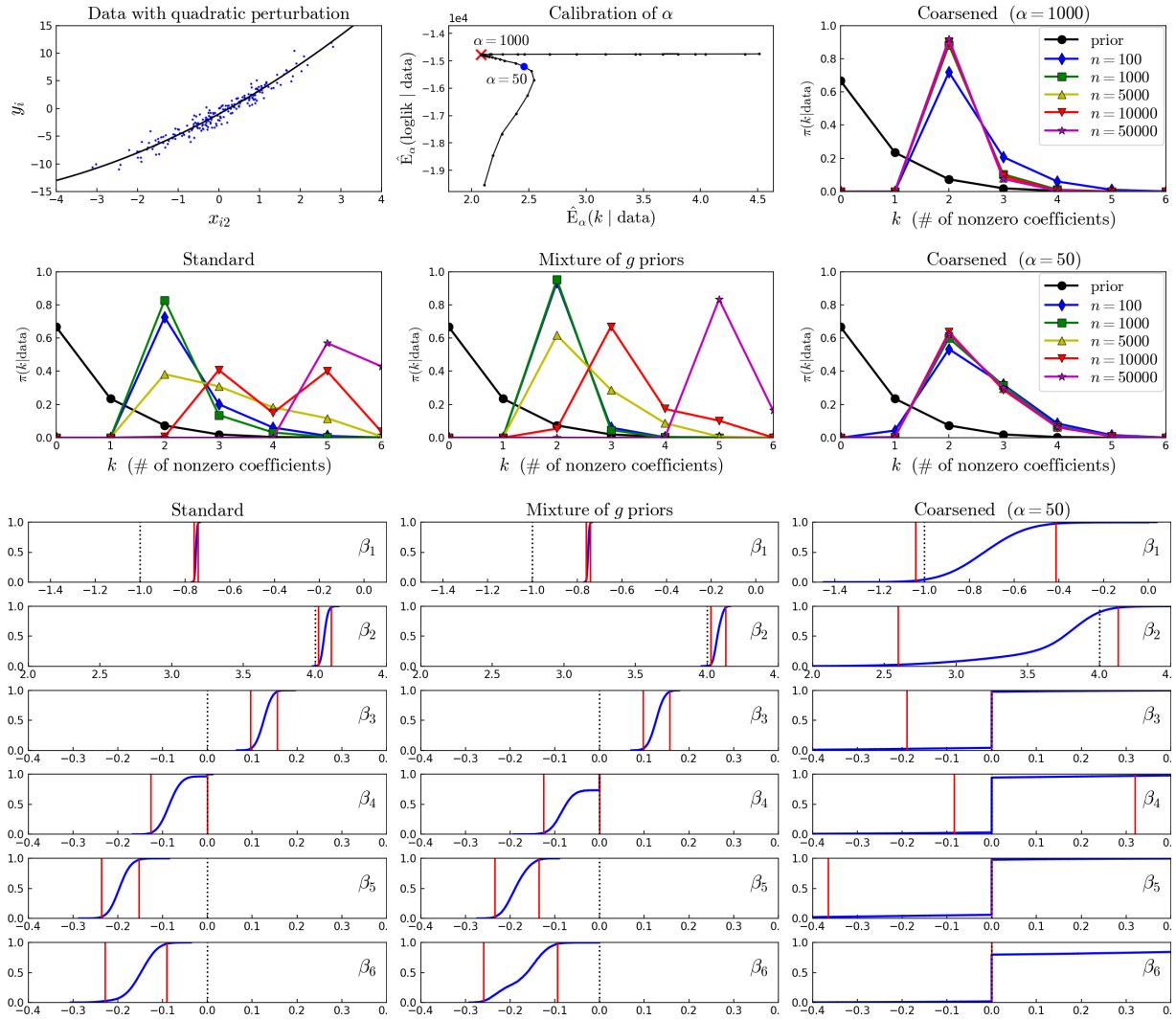


Figure 8: Variable selection on simulation with quadratic perturbation. Row 1 (left): Scatterplot of the target variable y_i versus x_{i2}, as well as the perturbed mean function (black line). Row 1 (middle): Calibration curve for α, when $n = 10000$. Rows 1 & 2 (right): The c-posteriors favor the true number of nonzero coefficients, $k = 2$, even as n grows. Row 2 (left): The standard posterior significantly overestimates k as n increases. Row 2 (middle): Using a robust prior (mixture of g priors) behaves very similarly to the standard posterior. Bottom: Posterior c.d.f.s for each coefficient (blue) and 95% credible intervals (red), when $n = 50000$; the true values are indicated by black dotted lines.

For the model parameters, we choose $r = 1$ and $s = 2p$ (in order to favor having $O(1)$ nonzero coefficients, regardless of p), $L_0 = 1$, and $a = b = 1$; this quantifies our prior beliefs about the true parameters. We describe two approaches to choosing α (which quantifies our trust in the likelihood), an *a priori* approach and a data-driven approach. For the *a priori* approach, note that the exact c-posterior is obtained by conditioning on the conditional relative entropy estimate being less than R, where $R \sim \text{Exp}(\alpha)$ (Supplement S6.2). The relative entropy between two Gaussians $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ is $\frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2$. Thus, if we expect the perturbation to shift the mean function by approximately $\pm\delta$ on average, and the noise has standard deviation σ, then it is natural to choose α so that $\mathbb{E}R \approx \delta^2/(2\sigma^2)$, i.e., $\alpha \approx 2\sigma^2/\delta^2$. In this example, by cheating and using our knowledge of the perturbation, we choose $\delta = 0.2$ and $\sigma = 1$, leading to $\alpha = 50$.

To choose α in a data-driven way, we can use the technique from Section 4. Figure 8 (top middle) shows the calibration curve on a dataset of size $n = 10^4$, with $f(\alpha) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \log p(y_{1:n}|\beta^{(t)}, \lambda^{(t)})$ to quantify fit (where $p(y_i|\beta, \lambda) = \mathcal{N}(y_i|\beta^\intercal x_i, 1/\lambda)$) and $g(\alpha) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} k(\beta^{(t)})$ to quantify effective complexity, where $k(\beta) = \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0)$ is the number of nonzero coefficients, and $\mathcal{T} = \{T_{\text{burn}} + 1, \ldots, T\}$. The curve exhibits a cusp with good fit and low complexity at $\alpha \approx 1000$, which suggests choosing $\alpha = 1000$.

For each $n \in \{100, 1000, 5000, 10000, 50000\}$, ten datasets were generated. For each of (i) the standard posterior, (ii) the posterior using a mixture of g priors, and (iii) the coarsened posterior with $\alpha = 50$ and $\alpha = 1000$, we ran Algorithm 7.1 for $T = 50000$ iterations on each dataset, discarding the first $T_{\text{burn}} = 5000$ iterations as burn-in.

For each method, Figure 8 shows the posteriors on k, averaged over the 10 datasets. In row 2, we see that using a mixture of g priors behaves essentially the same as the standard posterior—as n grows, both significantly overestimate k. Meanwhile, the c-posteriors with $\alpha = 50$ and $\alpha = 1000$ both favor the true value, $k = 2$ (see rows 1 & 2, right).

Figure 8 (bottom) shows the posterior cumulative distribution functions (c.d.f.s) and 95% credible intervals for each coefficient $\beta_1, \ldots, \beta_6$ when $n = 50000$. Recall that the true values are $\beta_1 = -1$, $\beta_2 = 4$, and $\beta_3 = \cdots = \beta_6 = 0$. The 95% intervals for both the

standard approach and mixture of g priors are quite far from the true values of β_1, β_3, β_5, and β_6. Meanwhile, all of the 95% intervals for the c-posterior contain the true values; also note that for $\beta_3, \ldots, \beta_6$, most of the c-posterior probability is at zero. The case of β_1, in particular, illustrates that the standard posterior can lead to incorrect inferences about the values of the nonzero coefficients, in addition to incorrectly inferring which coefficients are nonzero. The c-posterior mitigates this by more appropriately calibrating the amount of concentration, however, the price to be paid is that this can cause the c-posterior to be more diffuse than necessary; for instance, leading to overly wide intervals for β_2.

7.2 Comparison with a flexible nonlinear effects model

Using a more flexible model, such as a model allowing for nonlinear effects, does not solve the problem we seek to address with coarsening. To illustrate, we compare with the Bayesian Smoothing Spline ANOVA (BSSANOVA) (Reich et al., 2009) on two examples: (i) the quadratic perturbation example from Section 7.1, and (ii) a Gaussian process (GP) perturbation example. In the GP perturbation example, we suppose the true function is still $-1 + 4x_{i2}$, and the covariates x_i are still generated using the same skew-normal distribution as in Section 7.1. Then, to simulate a perturbation, we generate a Gaussian process $g_j \sim \text{GP}(0, \kappa_j)$ with covariance function $\kappa_j(z, z') = \sigma_j^2 \exp(-\frac{1}{2}(z - z')^2)$ for each $j = 2, \ldots, 6$, where $\sigma_2 = 1$ and $\sigma_3 = \cdots = \sigma_6 = 1/4$. The observed data are then generated as $y_i = -1 + 4x_{i2} + \sum_{j=2}^{6} g_j(x_{ij}) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$.

BSSANOVA models the data as $Y_i \sim \mathcal{N}(f(x_i), 1/\lambda)$ given f and λ, and places a non-parametric prior on the regression function: $f(x_i) = \mu + \sum_{j=2}^{p} f_j(x_{ij}) + f_0(x_{i2}, \ldots, x_{ip})$, where the f_j's and f_0 are GPs with carefully constructed covariance functions (see Reich et al., 2009). (Reich et al. (2009) present a more general version including two-way interaction terms $f_{jk}(x_{ij}, x_{ik})$, but interactions are not needed here.) A flat prior is placed on μ, and $\lambda \sim \text{Gamma}(a, b)$. Variable selection is implemented by placing a spike-and-slab prior on the scale of each f_j, so that $f_j \equiv 0$ with positive probability. We use the R code

for BSSANOVA provided by Reich et al. (2009), with the default settings for all model parameters. On each dataset, we run BSSANOVA for 10^4 MCMC iterations, discarding the first 1000 as burn-in.

Figure 9 shows the results. For the c-posterior, we use the same settings as in Section 7.1, with $\alpha = 50$. To assess the accuracy of inference for the true (unperturbed) function $f_{\theta_I}(x_i) = -1 + 4x_{i2}$, we compute 99% credible intervals for $f_{\theta_I}(x_i)$ under the standard posterior, c-posterior, and BSSANOVA posterior. In Figure 9 (left top/bottom), these intervals are shown (with one marker indicating each endpoint) for the first 200 data points, when $n = 5000$. BSSANOVA concentrates around the perturbed function, since this is what the observed data comes from. Meanwhile, the c-posterior provides more appropriately calibrated inference for the true function. Only 40% (quadratic example) and 26% (GP example) of the BSSANOVA intervals contain the true value $f_{\theta_I}(x_i)$, whereas 100% of the c-posterior intervals contain the true value; note that this fraction depends on the particular dataset and on average would be 99% across repeated datasets if the coverage were perfect.

For variable selection, BSSANOVA strongly favors the true value of $k = 2$ (x_{i1} and x_{i2}) in the quadratic example, as expected, since the perturbation is a function of x_{i2} only. However, in the GP example, BSSANOVA tends to overestimate the true k, using several additional covariates to fit the perturbation terms $g_j(x_{ij})$ for $j = 3, \ldots, 6$. Meanwhile, the c-posterior still favors the true value of $k = 2$ in the GP example.

This should in no way be interpreted as a criticism of BSSANOVA — it is doing exactly what it was designed to do: fit the observed data distribution. The problem is that inference for the true (pre-perturbation) parameters requires one to specifically consider perturbations from the idealized model, as we do in coarsening. Alternatively, instead of coarsening, one could augment the idealized model with a nonlinear perturbation model, but this is not the same as just using a flexible nonlinear model like BSSANOVA.

The computation time required is shown in Figure 9 (top right). BSSANOVA (implemented in R) takes roughly three orders of magnitude longer than the c-posterior (implemented in Julia). To compare on a more equal footing, language-wise, we also run

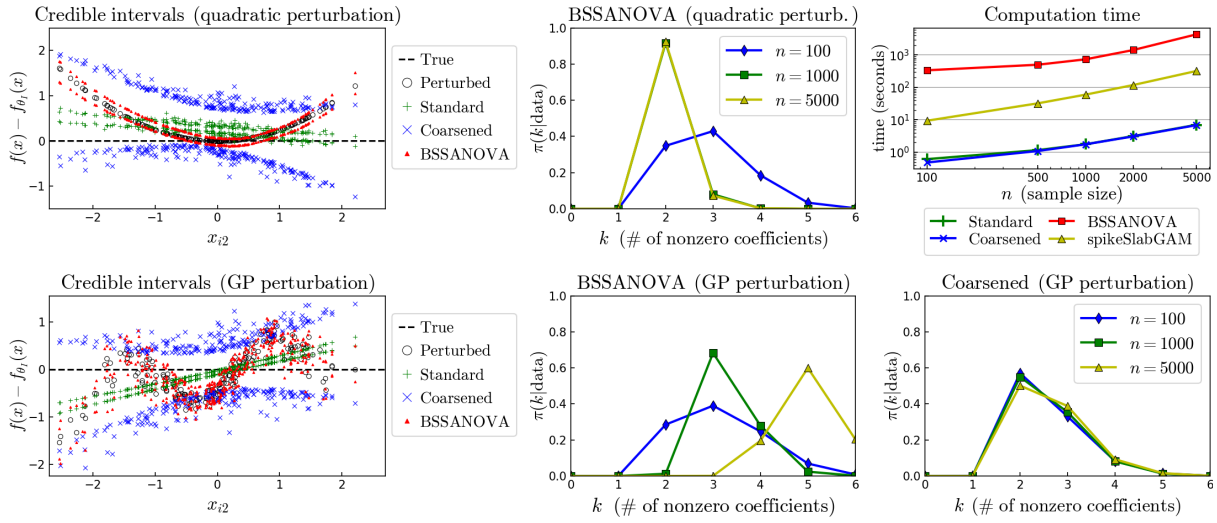


Figure 9: Comparison with a nonlinear effects model (BSSANOVA). Left: 99% credible intervals for $f_{\theta_I}(x_i)$ for each method, along with the true function f_{θ_I} and the perturbed function values. The plots are relative to f_{θ_I} for ease of visualization, and are projected onto x_{i2}. Middle: BSSANOVA strongly favors the true value of k for the quadratic perturbation, but overestimates k for the GP perturbation. Bottom right: The c-posterior still favors the true value of k. Top right: Computation time required for 10^4 MCMC iterations.

spikeSlabGAM (Scheipl et al., 2012), which is very similar to BSSANOVA but is implemented in C for speed. spikeSlabGAM is still around 1.5 orders of magnitude slower.

7.3 Application: Effect of chemical exposures on birth weight

During prenatal development, children are particularly susceptible to chemical exposures, making it important to understand the effect of common exposures on infant health outcomes (Fein et al., 1984; Wickerham et al., 2012). Estimating these effects from observational data is difficult since the causal structure among the variables is complex and unknown. Unobserved confounders or departures from assumed causal relationships can lead to spurious findings, especially with large sample sizes. To illustrate how coarsening can help diagnose and provide robustness to small departures from model assumptions, we consider the effect of various exposures on birth weight, using data from the Collaborative

Perinatal Project (Klebanoff, 2009), a large cohort study used for many purposes. Birth weight is a commonly used proxy for quality of gestation, since low birth weight infants more frequently experience a variety of health and developmental problems.

The dataset we use contains $n = 2379$ subjects and 45 predictor variables, including 31 exposures of interest (such as DDEs, PCBs, and smoking) and 14 control variables (such as race and gender). The data are preprocessed to normalize each predictor as well as the target variable (birth weight), by subtracting off the sample mean and dividing by the sample standard deviation for each. For each predictor, missing entries are imputed using the non-missing sample mean; for the target variable, there is one subject with a missing entry, which we exclude from the analysis. (It would be preferable, but more complex, to model the missing data.) A constant covariate is appended for the intercept, making $p = 46$. We use the same prior parameters as in Section 7.1, and for each run of Algorithm 7.1 we use $T = 20000$ iterations, with $T_{\text{burn}} = 2000$.

Following recommended practice (VanderWeele and Shpitser, 2011), we chose the 14 control variables by identifying (as best as possible based on prior knowledge) those variables that may have a causal effect on exposure or outcome, but are not causally affected by exposure. One variable that is ambiguous in this regard is mother's weight, prepregnancy (V_MWGTPP). We decided to exclude V_MWGTPP initially since, for example, smoking could reduce mother's weight, and reduced mother's weight could cause lower birth weight.

For the standard posterior, Figure 10 (bottom left) shows the posterior c.d.f.s of the coefficients for the top variables, i.e., the variables with the highest inclusion probability. Adjusting for control variables, this initial analysis suggests that cigarettes smoked per day (V_CPDNOW) and trans-Nonachlor (NONA_A) may decrease birth weight, and triglycerides (TRIGLYC) may increase it. Trans-Nonachlor is a component of the organochlorine pesticide chlordane, and the literature is not conclusive regarding the effect of these pesticides on birth weight (Gladen et al., 2003; Wickerham et al., 2012; Neta et al., 2011).

For the coarsened posterior, to choose α we plot the calibration curve in Figure 10 (top left). There is not a clear choice of α yielding good fit and low complexity, but the curve

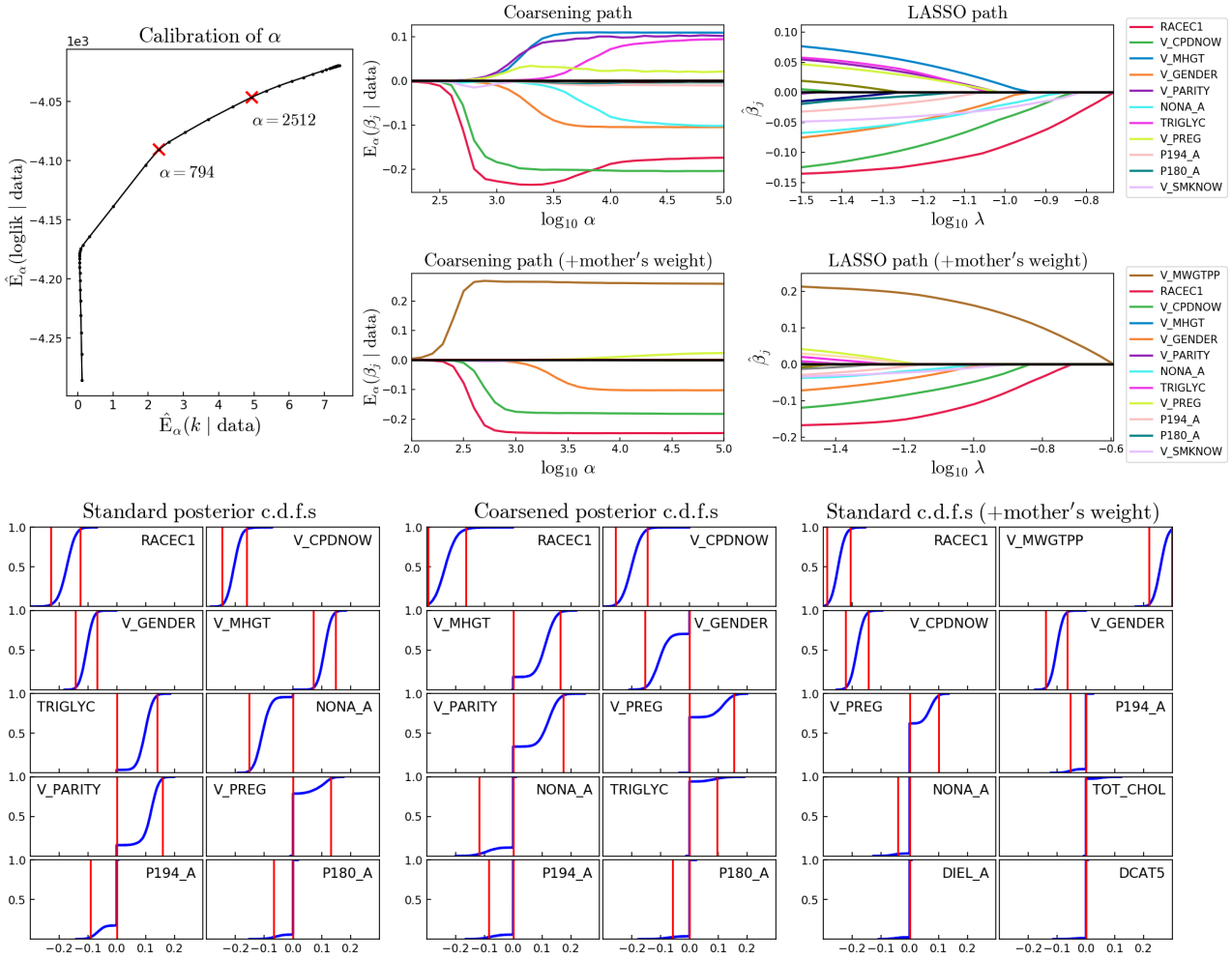


Figure 10: Effect of exposures on birth weight. Top left: Calibration curve for α (without mother's weight). Top right: Coarsening path and LASSO path, with and without mother's weight (V_MWGTPP) as a control variable. Bottom: Posterior c.d.f.s of the coefficients of the top variables, for the standard posterior without mother's weight (left), the c-posterior without mother's weight when $\alpha = 2500$ (middle), and the standard posterior with mother's weight (right). Variables shown: RACEC1: African-American; V_GENDER: Child's gender; V_MHGT: Maternal height; V_PREG: Ever pregnant; V_PARITY: Parity; V_MWGTPP: Maternal weight, pre-pregnancy; V_SMKNOW: Smoking status, now; V_CPDNOW: Cigarettes/day, now; DCAT5: DDE serum levels ≥ 60 ug/L; TOT_CHOL: Total cholesterol, ug/L; TRIGLYC: Triglycerides, ug/L; DIEL_A: Dieldrin, ug/L; NONA_A: trans-Nonachlor, ug/L; P180_A: PCB 180, ug/L; P194_A: PCB 194, ug/L.

suggests $\alpha \approx 800$ and $\alpha \approx 2500$ as candidates to explore, since they exhibit fair fit and are points where the curve changes slope. For $\alpha = 2500$, the posterior c.d.f.s are shown in Figure 10 (bottom middle). Compared to the standard posterior, we see that NONA_A and TRIGLYC no longer appear to have a strong effect.

This prompted us to consider other explanations for the apparent effect of NONA_A and TRIGLYC in the standard posterior. For instance, triglycerides and mother's weight are positively correlated (perhaps due to an unobserved confounder such as diet) and it seems likely that mother's weight affects birth weight; thus, the apparent effect of triglycerides could be a spurious consequence of omitting mother's weight. Similarly, mother's weight (V_MWGTPP) is negatively correlated with NONA_A and PCB 194 (P194_A), indicating that the apparent effect of NONA_A and P194_A might also be due to confounding.

To evaluate this possibility, we reran the analysis with V_MWGTPP added to the list of control variables. Figure 10 (bottom right) shows the coefficient c.d.f.s under the standard posterior with V_MWGTPP added as a covariate. This yields a more parsimonious explanation with V_MWGTPP taking the place of V_MHGT (mother's height), TRIGLYC, NONA_A, and V_PARITY, indicating that these other variables were previously included due to their association with V_MWGTPP. It seems likely that the apparent effect of NONA_A and TRIGLYC in the initial analysis was due to confounding. This example illustrates how the c-posterior helps deal with misspecification due to an omitted variable, by aiding in identifying the variable and providing some robustness in the results.

In this example, we also compare with the LASSO (Tibshirani, 1996). Figure 10 (top right) shows the LASSO path, that is, the coefficient estimates $\hat{\beta}_j$ for each value of the LASSO regularization parameter λ. Similarly, we can define a "coarsening path" by plotting the posterior means of the β_j's for each value of the coarsening parameter α; see Figure 10 (top middle). It is interesting that as α increases, the c-posterior estimates tend to be more stable once they are established, whereas the LASSO estimates tend to continually increase in magnitude as λ decreases. It appears that LASSO estimates are more sensitive to the precise choice of this parameter than c-posterior estimates.

In the LASSO, there is a duality between increasing the regularization parameter λ and the decreasing the weight given to the log likelihood, since $\operatorname{argmin}_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ is unaffected by multiplicative scaling. However, it is important to note that there is no such duality in the c-posterior. In the c-posterior, strengthening the prior is not equivalent to weakening the likelihood via α, because strengthening the prior does not make the posterior less concentrated (whereas weakening the likelihood does).

8 Toy example: Perturbed Normal with outliers

Suppose the true (idealized) distribution is known to be of the form $\mathcal{N}(\mu_0, \sigma_0^2)$, but the observed data $x_1, \ldots, x_n$ come from an unknown nonparametric perturbation of $\mathcal{N}(\mu_0, \sigma_0^2)$ with a large fraction of outliers (see Figure 11, top left).

To obtain a c-posterior that is robust to outliers, we use the 1st Wasserstein distance as our choice of discrepancy. For empirical distributions on n points in $\mathbb{R}$, the 1st Wasserstein distance simplifies to $d_n(x'_{1:n}, x_{1:n}) = \frac{1}{n}\sum_{i=1}^n |x'_{(i)} - x_{(i)}|$, where $x_{(i)}$ is the ith order statistic. We define the model as $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ since the true distribution is normal, and we place priors on the parameters: $\mu \sim \mathcal{N}(m, s^2)$ and $\sigma^2 \sim \text{InverseGamma}(a, b)$, independently. We place an $\text{Exp}(\alpha)$ prior on the coarsening radius R, as before. To infer α, we can use a Bayesian approach since $d_n(\cdot,\cdot)$ is tractable and the perturbation is not approximable by the model. Thus, we place a $\text{Gamma}(u, v)$ prior on α. We perform inference for the resulting Wasserstein c-posterior using the following MCMC algorithm to sample from $\pi\big(\theta, \alpha \mid d_n(X_{1:n}, x_{1:n}) < R\big)$ where $\theta = (\mu, \sigma^2)$; see Supplement S7.3 for the derivation. Define $\pi(z_{1:n}) = \prod_{i=1}^n \mathcal{N}(z_i|0,1)$ and $w_{\theta,\alpha}(z_{1:n}) = \exp\big(-\alpha d_n(h_\theta(z_{1:n}), x_{1:n})\big)$, where $h_{(\mu,\sigma^2)}(z_{1:n}) = (z_1\sigma + \mu, \ldots, z_n\sigma + \mu)$. Note that $\pi(\theta) = \mathcal{N}(\mu|m, s^2)\text{InverseGamma}(\sigma^2|a, b)$.

Algorithm 8.1 (MCMC sampler for univariate normal Wasserstein c-posterior)**.**

- *Input:* $x_1, \ldots, x_n$. *Output: MCMC samples of* $\theta = (\mu, \sigma^2)$ *and* α.
- *Initialize* $\theta = (0, 1)$, $\alpha = 1$, *and* $z_1, \ldots, z_n \sim \mathcal{N}(0, 1)$.
- *For iteration* $t = 1, \ldots, T$*:*

1. *Update* α *by sampling* $\alpha \sim \text{Gamma}\big(u,\, v + d_n(h_\theta(z_{1:n}), x_{1:n})\big)$.
2. *Propose* $\theta' \sim \mathcal{N}(\theta, \varepsilon_\theta^2 I)$ *and accept with probability* $\min\left\{1, \dfrac{w_{\theta',\alpha}(z_{1:n})\pi(\theta')}{w_{\theta,\alpha}(z_{1:n})\pi(\theta)}\right\}$.
3. *Propose* $z'_{1:n} \sim \mathcal{N}(z_{1:n}, \varepsilon_z^2 I)$ *and accept with probability* $\min\left\{1, \dfrac{w_{\theta,\alpha}(z'_{1:n})\pi(z'_{1:n})}{w_{\theta,\alpha}(z_{1:n})\pi(z_{1:n})}\right\}$.

To demonstrate, suppose the true parameters are $\mu_0 = 3.2$ and $\sigma_0^2 = 4.4$. To simulate a perturbation of $\mathcal{N}(\mu_0, \sigma_0^2)$ with outliers, we generate the observed x_i's by sampling i.i.d. from $0.9P + 0.1\mathcal{N}(20, 1)$, where P is a random draw of a Dirichlet process mixture with base distribution $\mathcal{N}(\mu_0, \sigma_0^2)$, concentration parameter 500, and $\mathcal{N}(0, 0.25^2)$ components. The perturbation is depicted in Figure 11 (top left). For the model hyperparameters, we set $m = 0$, $s = 5$, $a = 1$, $b = 1$, $u = 7$, and $v = 0$; this results in an improper prior on α that enables the c-posterior to concentrate if the model is correct (see Supplement S7.3 for details). For the proposal distributions, we use $\varepsilon_\theta = 0.25$ and $\varepsilon_z = 0.02$. For each run of Algorithm 8.1, we use $T = 2 \times 10^5$ iterations, discarding the first 10% as burn-in.

Figure 11 shows the standard posterior and the c-posterior on perturbed datasets of sizes $n \in \{50, 200, 1000, 10000\}$. Due to the perturbation, the standard posterior concentrates at values of μ and σ^2 that are quite far from the true values, $\mu_0 = 3.2$ and $\sigma_0^2 = 4.4$ (black dotted lines). Meanwhile, the c-posterior is appropriately calibrated for all values of n.

In Figure 12, we compare results when there is no perturbation (and thus, the Normal model is correct), by simulating the observed data as $x_1, \ldots, x_n$ i.i.d. $\sim \mathcal{N}(\mu_0, \sigma_0^2)$ and running the same posterior inference algorithms. In this case, both the standard posterior and the c-posterior appear to be concentrating at the true values. Note that the c-posterior favors increasingly large values of α as $n \to \infty$, since it adaptively infers that little or no coarsening is required on the unperturbed data.

9 Conclusion

The c-posterior approach seems promising as a general method of robust Bayesian inference. There are several directions that would be interesting to pursue in future work. Further

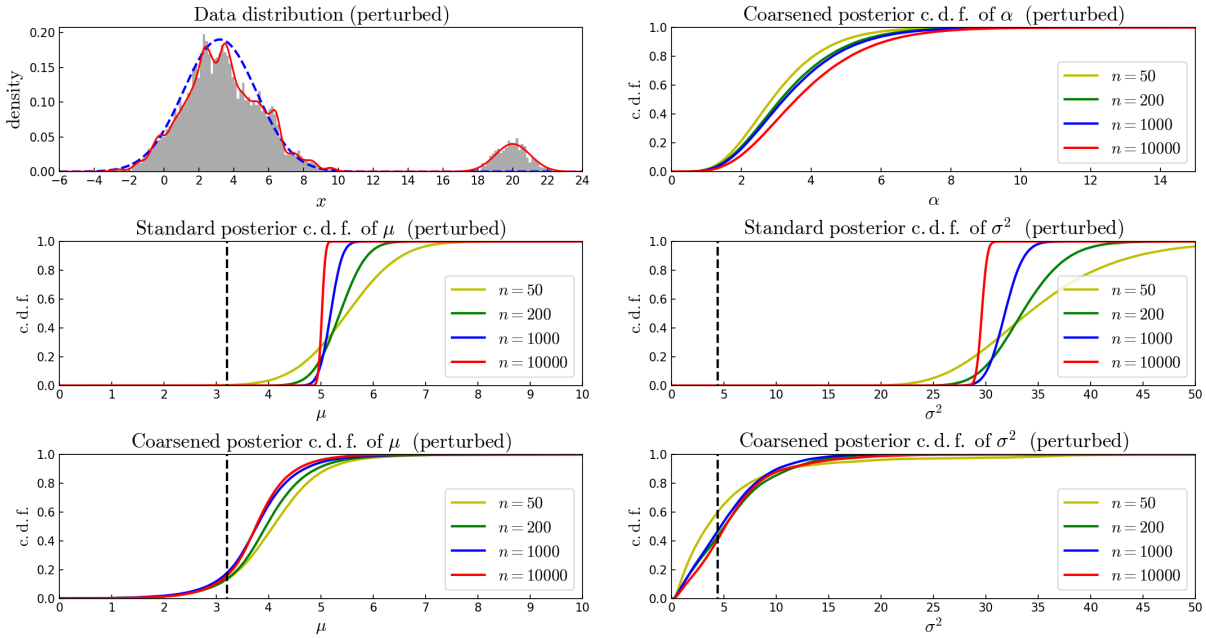


Figure 11: Perturbed Normal example. Top left: True density (blue dotted line), perturbed density (red line), and histogram of data. Top right: Posterior on α. Middle/bottom: The standard posterior concentrates at incorrect values, whereas the c-posterior is appropriately calibrated.

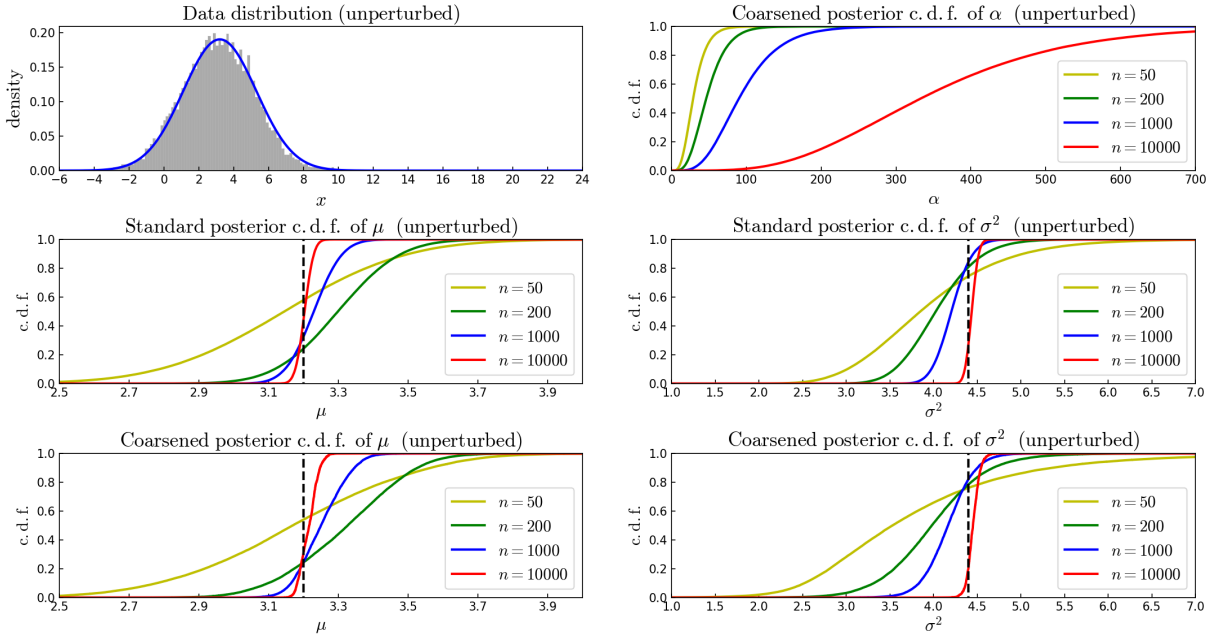


Figure 12: Normal example without perturbation. Top left: True density (blue line) and histogram of unperturbed data. Top right: Posterior on α. Middle/bottom: Both the standard posterior and the c-posterior appear to be concentrating at the true values of μ and σ^2, at similar rates.

investigation of the accuracy of the power posterior approximation is needed, both theoretically and empirically. Additionally, it would be beneficial if precise guarantees could be provided regarding frequentist coverage properties of the c-posterior when there is a perturbation. Finally, it would be interesting to explore coarsening in frequentist procedures, since the scope of application is not limited to Bayesian inference.

Acknowledgments

We would like to thank the editors and referees for many helpful suggestions that improved the quality of this manuscript. We would also like to thank Matthew Harrison, Stuart Geman, Erik Sudderth, Jacopo Soriano, Peter Grünwald, and Mark Glickman for insightful conversations.

References

Aghaeepour, N., Nikolic, R., Hoos, H. H., and Brinkman, R. R. Rapid cell population identification in flow cytometry data. *Cytometry Part A*, 79(1):6–13, 2011.

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R., Gottardo, R., Scheuermann, R. H., FlowCAP Consortium, DREAM Consortium, and 91 others. Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238, 2013.

Antoniano-Villalobos, I. and Walker, S. G. Bayesian nonparametric inference for the power likelihood. *Journal of Computational and Graphical Statistics*, 22(4):801–813, 2013.

Azzalini, A. and Capitanio, A. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577, 2012.

Brinkman, R. R., Gasparetto, M., Lee, S.-J. J., Ribickas, A. J., Perkins, J., Janssen, W., Smiley, R., and Smith, C. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, 13(6):691–700, 2007.

Fein, G. G., Jacobson, J. L., Jacobson, S. W., Schwartz, P. M., and Dowler, J. K. Prenatal exposure to polychlorinated biphenyls: effects on birth size and gestational age. *The Journal of Pediatrics*, 105(2):315–320, 1984.

Finak, G., Bashashati, A., Brinkman, R., and Gottardo, R. Merging mixture components for cell population identification in flow cytometry. *Advances in Bioinformatics*, 2009 (247646), 2009.

Gladen, B. C., Shkiryak-Nyzhnyk, Z. A., Chyslovska, N., Zadorozhnaja, T. D., and Little, R. E. Persistent organochlorine compounds and birth weight. *Annals of Epidemiology*, 13(3):151–157, 2003.

Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2006.

Hoff, P. D. *A First Course in Bayesian Statistical Methods*. Springer Science & Business Media, 2009.

Ishwaran, H. and Zarepour, M. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963, 2002.

Klebanoff, M. A. The Collaborative Perinatal Project: a 50-year retrospective. *Paediatric and Perinatal Epidemiology*, 23(1):2–8, 2009.

Neta, G., Goldman, L. R., Barr, D., Apelberg, B. J., Witter, F. R., and Halden, R. U. Fetal exposure to chlordane and permethrin mixtures in relation to inflammatory cytokines and birth outcomes. *Environmental Science & Technology*, 45(4):1680–1687, 2011.

Reich, B. J., Storlie, C. B., and Bondell, H. D. Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics*, 51 (2):110–120, 2009.

Scheipl, F., Fahrmeir, L., and Kneib, T. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 267–288, 1996.

VanderWeele, T. J. and Shpitser, I. A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413, 2011.

Wickerham, E. L., Lozoff, B., Shao, J., Kaciroti, N., Xia, Y., and Meeker, J. D. Reduced birth weight in relation to pesticide mixtures detected in cord blood of full-term infants. *Environment International*, 47:80–85, 2012.

Supplementary material for "Robust Bayesian inference via coarsening"

S1 Previous work

The c-posterior is mathematically equivalent to the posterior approximation resulting from approximate Bayesian computation (ABC) (Tavaré et al., 1997; Marjoram et al., 2003; Beaumont et al., 2002; Wilkinson, 2013). However, our motivation is completely different from that of ABC—we are concerned with robustness to perturbations, whereas ABC is concerned with inference in models with intractable likelihoods. Generally speaking, we assume the likelihood can be computed easily, which makes our inferences much more computationally efficient.

The c-posterior can also be viewed as conditioning on partial information, a technique that is often used to improve robustness (Doksum and Lo, 1990; Pettitt, 1983; Hoff, 2007; Dunson and Taylor, 2005; Lewis et al., 2014); also see Cox (1975). Usually, however, this is done by conditioning on some insufficient statistic; for example, Doksum and Lo (1990) perform robust Bayesian inference for a location parameter by conditioning only on the sample median, rather than the whole sample. Our approach of conditioning on a distributional neighborhood is quite different.

Gibbs posteriors have recently been introduced as a general framework for updating prior beliefs using a generalized "likelihood" (Jiang and Tanner, 2008; Zhang, 2006b; Li et al., 2014; Holmes et al., 2016). Under certain conditions (see Supplement S3.1), when n is large the c-posterior is approximately proportional to $\exp(-\alpha d(P_\theta, \hat{P}_{x_{1:n}}))\pi(\theta)$, which can be viewed as a Gibbs posterior with "risk" $d(P_\theta, \hat{P}_{x_{1:n}})$. In research involving Gibbs posteriors, an issue of current interest is how to choose α so that the concentration of the posterior is appropriately calibrated. The connection between Gibbs posteriors and c-posteriors may provide insight into this calibration problem.

A number of researchers have used the technique of raising the likelihood to a fractional power. Usually, this is done for reasons completely unrelated to robustness, such as marginal likelihood approximation (Friel and Pettitt, 2008), improved MCMC mixing (Geyer, 1991), consistency in nonparametric models (Walker and Hjort, 2001; Zhang, 2006a), discounting historical data (Ibrahim and Chen, 2000; Smith, 1981), or objective Bayesian model selection (O'Hagan, 1995). However, recently, the robustness properties of power likelihoods have been noticed: Grünwald and van Ommen (2014) provide an in-depth study of a simulation example in which a power posterior exhibits improved robustness to misspecification, and they propose a method for choosing the power; also see Royall and Tsou (2003) and Ghosh and Sudderth (2012). In all such previous work, a fixed power is used, rather than a power tending to 0 as $n \to \infty$. It seems that neither the form of power likelihood we use, nor the theoretical motivation for it, have appeared in any prior work.

Most of the previous work on Bayesian robustness has been concerned with robustness to the choice of prior, rather than robustness to the form of the likelihood. Robustness to the prior is often formulated from a decision-theoretic perspective in which one chooses a decision rule that minimizes the worst-case Bayes risk over some set of priors; this is known as the Γ-minimax approach (Berger, 1985; Berger and Berliner, 1986). In a similar vein, minimax decision-theoretic approaches for robustness to the likelihood have also been explored: Hansen and Sargent (2001) propose choosing a decision rule that minimizes worst-case expected loss over a set of data distributions within a neighborhood of some point estimate; also see Whittle (1990) and Watson and Holmes (2016). These decision-theoretic approaches are appealing, but are quite different from what we propose.

Conceptually, the existing methods that seem most similar to the idea of the c-posterior are goodness-of-fit tests that assess whether the data distribution is close to the set of model distributions (Rudas et al., 1994; Goutis and Robert, 1998; Carota et al., 1996; Dette and Munk, 2003; Liu and Lindsay, 2009), however, the methods used previously are very different from ours. Related to such work is the model credibility index of Lindsay and Liu (2009), which has heavily influenced our thinking in the development of the c-posterior.

S2 Discussion

The c-posterior approach has a number of appealing features. It has a compelling justification—it is valid Bayesian inference based on limited information. The interpretation is conceptually clear—one does inference with the same model, but conditioned on a different event than usual. As shown in Section S3, the c-posterior inherits the continuity properties of the chosen discrepancy, and thus, exhibits robustness to small perturbations. Below, we address several frequently asked questions.

Concentration versus calibration

The main disadvantage of the c-posterior is that sometimes it is less concentrated than one would like. This problem is illustrated by the overly wide c-posterior credible interval for β_2 in the variable selection example shown in Figure 8. On the other hand, the same example shows that when there is a perturbation, c-posterior credible intervals tend to behave better than standard credible intervals. Less concentration (wider intervals) is the price to be paid to obtain a better calibrated posterior in the sense of frequentist coverage.

Not equivalent to renormalized tempering or overdispersion

We would like to emphasize that the power posterior is not equivalent to the posterior under a model with density $f(x|\theta,\zeta) \propto p_\theta(x)^\zeta$ (where $\propto$ indicates proportionality with respect to x) since the normalization constant of f involves θ, whereas the power likelihood does not contain this normalization constant. Using a model based on f would not be expected to provide the same robustness properties as the power posterior, since it simply amounts to a model with one additional parameter, ζ.

Measurement error

A frequently asked question is whether the problems addressed by the c-posterior could instead be handled using measurement error methods.

The term "measurement error" usually refers to the situation in which the covariates in a regression model are observed with error (Carroll et al., 2006). This represents one particular kind of perturbation, and it is usually dealt with by changing the model appropriately in order to make it correctly specified. We are concerned with the broader class of misspecification problems in general — not just covariate error, and not just regression models. Further, in many situations it is impractical to correct the model, and these are the situations our method is intended to address.

Alternatively, sometimes "measurement error" is used to refer to an augmentation of the model to account for additional error/noise/uncertainty in the observed data, beyond what is already included in the original model. There are essentially two ways of doing this, the first of which does not solve the fundamental problem addressed by the c-posterior, and the second of which tends to be computationally expensive:

1. One could assume a model for the distribution of $x_i|X_i$ (in the notation of Section 2), for example, Gaussian or some other error distribution. However, this simply amounts to convolving the original model distribution P_θ with the chosen error distribution, leading to a new model that has a few more parameters but is just as bound to be misspecified as the original model. For instance, if one is using a Gaussian mixture model, and then introduces an additional Gaussian error distribution for $x_i|X_i$, the result is simply a new Gaussian mixture model with inflated variances, which is likely to suffer from the same misspecification issues as the original model. Even if one nonparametrically models the error distribution for $x_i|X_i$, this is still more restrictive than our approach of allowing for a distributional perturbation from the original model, rather than just a convolution of it.

2. The second approach would be to jointly model the distribution of $x_{1:n}|X_{1:n}$. In principle, this can work, but the choice of distribution for $x_{1:n}|X_{1:n}$ cannot be something simple, otherwise this ends up suffering from the same issue as in item 1. In order for this approach to work well, the distribution of $x_{1:n}|X_{1:n}$ needs to allow for distri-

butional perturbations even as $n \to \infty$; essentially, it needs to be a nonparametric model for the empirical distribution $\hat{P}_{x_{1:n}}$ given $\hat{P}_{X_{1:n}}$ (see Figure 1). But this seems just as computationally burdensome as using a nonparametric model for P_o given P_{θ_I}, and then modeling $x_1, \ldots, x_n$ as i.i.d. from P_o. The point of our approach is that it behaves similarly to using such a nonparametric model, but is computationally far more efficient.

Separation of the amount of coarsening from the choice of prior

A question that may be asked is whether there is a duality between coarsening and using a robust prior. In particular, would less coarsening be required if one used a more robust prior? The answer is "no" for two reasons, one technical and one conceptual.

The technical reason is that the likelihood overwhelms the prior as the sample size increases. Thus, no matter what prior is selected, a perturbation involving the likelihood will require the same amount of coarsening. A robust prior provides robustness to the choice of prior, but not robustness to the choice of likelihood. For example, in Figure 8, we see that using a mixture of g priors (a leading example of a robust prior) yields results nearly identical to using our original choice of prior. Coarsening addresses the problem by dealing with the likelihood directly.

Confusion over a perceived duality may arise due to a misplaced analogy with penalized regression methods such as LASSO. In methods such as LASSO, there is a duality between the regularization coefficient of the penalty term and the weight given to the log likelihood term, since both terms can be multiplied by a constant without affecting the optimizing value. In contrast, when constructing a posterior distribution rather than computing an optimum, the concentration of the distribution is affected if one multiplies both terms (in this case, the log likelihood and the log prior) by a constant. Thus, adjusting the strength of the prior is not equivalent to adjusting the strength of the likelihood. The analogy does hold in one respect, which is that reducing the strength of *likelihood* (as in coarsening) is akin to increasing the regularization coefficient in penalized regression. This is illustrated

by the comparison between the LASSO path and the "coarsening path" in Figure 10.

There is also a conceptual reason for separation between the choice of prior and the amount of coarsening. In many applications, the parameters represent a true state of nature that has a meaning and existence completely separate from any likelihood. The prior represents our prior beliefs about this true state, regardless of any data generating mechanism or misspecification thereof. For instance, suppose θ is the height of a particular person. The prior distribution represents our uncertainty in θ before we have any idea what type of data may be received, and thus, before any likelihood is specified. Then, various data on the person's height may be received — such as self-reported height, measurement with a scale, parents' heights, or estimation from a photograph — which can be used to form a posterior by assuming a likelihood. The amount of coarsening required pertains to the amount of misspecification of the assumed likelihood (or equivalently, the magnitude of the perturbation), which is completely unrelated to the prior beliefs.

Strategies for choosing the amount of coarsening

There are several possible strategies for choosing α.

- *Strategy #1: Calibration curve.* Plot $(g(\alpha), f(\alpha))$ as α ranges from 0 to ∞, where $f(\alpha)$ is a measure of fit to the data and $g(\alpha)$ is a measure of effective complexity; then choose α at a point where the curve achieves good fit at low complexity; see Section 4 for details. This strategy is illustrated in Sections 3, 5.1, 6, 7.1, and 7.3.

- *Strategy #2: A priori knowledge.* Set the mean neighborhood size $\mathbb{E}R = 1/\alpha$ according to the expected size of the perturbation. To help quantify *a priori* knowledge in terms of neighborhood size R, it is possible in some cases to roughly translate intuitive notions like Euclidean distance into relative entropy. This strategy is illustrated in Sections 3 and 7.1.

- *Strategy #3: Inference.* If the chosen discrepancy $d_n(X_{1:n}, x_{1:n})$ can be easily computed, then one can attempt to infer α by comparing the observed data $x_{1:n}$ to

idealized data $X_{1:n}$ generated by an inferred model. This strategy is illustrated in Section 8. This strategy is best suited to situations in which the model cannot approximate the perturbation. The reason we do not apply this strategy in the other examples is that the relative entropy is difficult to estimate for continuous distributions due to the $\int p_o \log p_o$ term, which requires a density estimate; this is computationally and statistically inefficient except in very low dimensional cases.

- *Strategy #4: Sensitivity analysis.* Consider a range of α values, for sensitivity analysis or exploratory analysis.

In Strategy #2, a useful rule of thumb is to set $\alpha = N$ in order to be robust to perturbations that would require at least N samples to distinguish. Recall that the power posterior can be interpreted as adjusting the sample size from n to $n\zeta_n$, in terms of concentration of the posterior. Thus, since $n\zeta_n \to \alpha$ as $n \to \infty$, choosing $\alpha = N$ can be interpreted as saying that we want the posterior to be only as concentrated as when N samples or fewer are available.

What type of deviations does coarsening tolerate?

The type of deviations tolerated by the c-posterior depends on the choice of discrepancy $d_n(\cdot, \cdot)$ between distributions, and the size of deviation tolerated is governed by the prior on R. For instance, choosing $d_n(\cdot, \cdot)$ to be Wasserstein distance yields a c-posterior that is robust to deviations from the model that are small in Wasserstein distance — i.e., any perturbation that is small in Wasserstein distance results in a small change to the c-posterior. Meanwhile, if we choose relative entropy (i.e., Kullback–Leibler divergence), then any perturbation that is small in Kullback–Leibler divergence results in a small change to the c-posterior.

For the simulations, we used examples in which the perturbation is small with respect to the chosen discrepancy, in order to illustrate robustness of the c-posterior to such perturbations. One could choose any perturbation that is small with respect to a given discrepancy,

and the corresponding c-posterior would be robust to it (Section S3.3).

Bayesian updating

It is important to note that c-posteriors do not follow the standard rule for Bayesian updating—that is, if one multiplies the c-posterior for a subset of the data times the c-likelihood for the rest of the data, this is not proportional to the c-posterior for the whole data set, in general. Interestingly, however, there is a more general rule for rational Bayesian belief revision, known as Jeffrey conditionalization (Diaconis and Zabell, 1982; Jeffrey, 1965; Joyce, 2008). Jeffrey conditionalization handles cases in which one is only given partial information, which is precisely the situation dealt with by the c-posterior.

Likelihood principle

A potential philosophical criticism of c-posteriors is that they do not, in general, adhere to the likelihood principle. However, many important statistical methods violate the likelihood principle, so the practical relevance of this point is dubious. Curiously, the power posterior does adhere to the likelihood principle.

Data augmentation issues

Generally speaking, it is straightforward to use MCMC for sampling from the power posterior. However, there is a subtle point that should be carefully observed. Often, latent variables are introduced into an MCMC scheme in order to facilitate moves or to improve mixing, and sometimes, such latent variables do not work in the same way for the power posterior. For example, in a mixture model, say, $\sum_{i=1}^{K} w_i f_{\varphi_i}(x)$, latent variables $z_1, \ldots, z_n$ indicating which component each datapoint comes from are often introduced so that the full conditional distributions for w, φ, and z take nice and simple forms. However, when using the power posterior, the likelihood is $\prod_{j=1}^{n} \big(\sum_{i=1}^{K} w_i f_{\varphi_i}(x_j) \big)^{\zeta_n}$, and it seems that introducing $z_1, \ldots, z_n$ no longer leads to nice full conditionals. Algorithm 5.1 is designed to approximate the power posterior via an algorithm that closely resembles the standard data

augmentation approach involving $z_1, \ldots, z_n$. Alternatively, in some cases it may be possible to use a different set of latent variables; see Antoniano-Villalobos and Walker (2013) for the case of mixtures.

S3 Theory

In this section, we establish the asymptotic form of c-posteriors as $n \to \infty$ (Section S3.1), the limit as the distribution of R converges to 0, with n fixed (Section S3.2), and the robustness properties of c-posteriors (Section S3.3). Let $\mathcal{X}$ and Θ be standard Borel spaces, and let $\mathcal{M}$ denote the space of probability measures on $\mathcal{X}$, equipped with the weak topology. Let $\{P_\theta : \theta \in \Theta\} \subseteq \mathcal{M}$ be a family of probability measures on $\mathcal{X}$ such that $\theta \mapsto P_\theta(A)$ is measurable for all measurable subsets $A \subseteq \mathcal{X}$. Let Π be a prior measure on Θ, and let

$$\begin{aligned} &\boldsymbol{\theta} \sim \Pi, \\ &X_1, \ldots, X_n | \boldsymbol{\theta} \text{ i.i.d.} \sim P_{\boldsymbol{\theta}}, \text{ and} \\ &R \sim H, \text{ independently of } \boldsymbol{\theta}, X_{1:n}, \end{aligned}$$

where H is a distribution on $[0, \infty)$. We use (bold) $\boldsymbol{\theta}$ for the random variable, and θ for particular values. Define $G(r) = \mathbb{P}(R > r)$. Suppose the observed data $x_1, \ldots, x_n \in \mathcal{X}$ behave like i.i.d. samples from some $P_o \in \mathcal{M}$. Let $d : \mathcal{M} \times \mathcal{M} \to [0, \infty]$, and for $n \in \{1, 2, \ldots\}$, let $d_n : \mathcal{X}^n \times \mathcal{X}^n \to [0, \infty]$. It is assumed that $\theta \mapsto d(P_\theta, P)$ is measurable for all $P \in \mathcal{M}$, and $d_n(\cdot, \cdot)$ is measurable for each n.

S3.1 Large-sample asymptotics of the c-posterior

The c-posterior takes a simple form as $n \to \infty$, under mild regularity conditions. The following basic lemma captures the underlying principle at work in establishing both the asymptotic form of the c-posterior (Theorem S3.3) and its robustness (Theorem S3.8).

Lemma S3.1. *If $U, U_n, V, W \in \mathbb{R} \cup \{\infty\}$ are random variables such that $U_n \xrightarrow[n\to\infty]{\text{a.s.}} U$, $\mathbb{P}(U = V) = 0$, $\mathbb{P}(U < V) > 0$, and $\mathbb{E}|W| < \infty$, then $\mathbb{E}(W \mid U_n < V) \xrightarrow[n\to\infty]{} \mathbb{E}(W \mid U < V)$.*

Proof. Since $\mathbb{P}(U = V) = 0$, we have $\mathbb{1}(U_n < V) \xrightarrow{\text{a.s.}} \mathbb{1}(U < V)$, and thus, also $W\mathbb{1}(U_n < V) \xrightarrow{\text{a.s.}} W\mathbb{1}(U < V)$. Hence, by the dominated convergence theorem (Breiman, 1968, 2.44), $\mathbb{P}(U_n < V) \longrightarrow \mathbb{P}(U < V)$ and

$$\mathbb{E}\big(W\mathbb{1}(U_n < V)\big) \longrightarrow \mathbb{E}\big(W\mathbb{1}(U < V)\big)$$

since $0 \leq \mathbb{1}(\cdot) \leq 1$, $|W\mathbb{1}(U_n < V)| \leq |W|$, and $\mathbb{E}|W| < \infty$. By assumption, $\mathbb{P}(U < V) > 0$, hence $\mathbb{P}(U_n < V) > 0$ for all n sufficiently large, and

$$\mathbb{E}(W|U_n < V) = \frac{\mathbb{E}\big(W\mathbb{1}(U_n < V)\big)}{\mathbb{P}(U_n < V)} \xrightarrow[n\to\infty]{} \frac{\mathbb{E}\big(W\mathbb{1}(U < V)\big)}{\mathbb{P}(U < V)} = \mathbb{E}(W|U < V).$$

□

The following condition is necessary to avoid certain pathologies; it is always satisfied when $d(P_{\boldsymbol{\theta}}, P_o) < \infty$ with positive probability and R has a density with respect to Lebesgue measure that is positive on $[0, \infty)$, for instance. We use $\Rightarrow$ to denote convergence with respect to the weak topology.

Condition S3.2. *Assume* $\mathbb{P}(d(P_{\boldsymbol{\theta}}, P_o) = R) = 0$ *and* $\mathbb{P}(d(P_{\boldsymbol{\theta}}, P_o) < R) > 0$.

Theorem S3.3. *If* $d_n(X_{1:n}, x_{1:n}) \xrightarrow[n\to\infty]{\text{a.s.}} d(P_{\boldsymbol{\theta}}, P_o)$ *and Condition S3.2 is satisfied, then*

$$\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) \xRightarrow[n\to\infty]{} \Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big) \propto G\big(d(P_\theta, P_o)\big)\Pi(d\theta), \quad \text{(S3.1)}$$

and in fact,

$$\mathbb{E}\big(h(\boldsymbol{\theta}) \mid d_n(X_{1:n}, x_{1:n}) < R\big) \xrightarrow[n\to\infty]{} \mathbb{E}\big(h(\boldsymbol{\theta}) \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big) = \frac{\mathbb{E}h(\boldsymbol{\theta})G\big(d(P_{\boldsymbol{\theta}}, P_o)\big)}{\mathbb{E}G\big(d(P_{\boldsymbol{\theta}}, P_o)\big)} \quad \text{(S3.2)}$$

for any $h \in L^1(\Pi)$*, i.e., any measurable* $h : \Theta \to \mathbb{R}$ *such that* $\int |h(\theta)|\Pi(d\theta) < \infty$.

Proof. We apply Lemma S3.1 with $U = d(P_{\boldsymbol{\theta}}, P_o)$, $U_n = d_n(X_{1:n}, x_{1:n})$, $V = R$, and $W = h(\boldsymbol{\theta})$. By assumption, $U_n \xrightarrow{\text{a.s.}} U$, $\mathbb{P}(U = V) = 0$, $\mathbb{P}(U < V) > 0$, and $\mathbb{E}|W| < \infty$. Hence, by Lemma S3.1,

$$\mathbb{E}(W \mid U_n < V) \longrightarrow \mathbb{E}(W \mid U < V) = \frac{\mathbb{E}(W\mathbb{1}(U < V))}{\mathbb{P}(U < V)}$$

$$= \frac{\mathbb{E}\big(W\mathbb{E}(\mathbb{1}(U < V)|W, U)\big)}{\mathbb{E}\big(\mathbb{P}(U < V \mid U)\big)} = \frac{\mathbb{E}(WG(U))}{\mathbb{E}G(U)}$$

since $V \perp U, W$ by construction. This establishes Equation S3.2, and since in particular this holds for any bounded continuous h, Equation S3.1 follows. □

A case of particular interest arises when $R \sim \text{Exp}(\alpha)$, since then $G(r) = e^{-\alpha r}$ and the resulting asymptotic c-posterior is proportional to $\exp(-\alpha d(P_\theta, P_o))\Pi(d\theta)$, by Theorem S3.3. This is asymptotically equivalent to $\exp(-\alpha d(P_\theta, \hat{P}_{x_{1:n}}))\Pi(d\theta)$, provided that $d(P_{\boldsymbol{\theta}}, \hat{P}_{x_{1:n}}) \xrightarrow{\text{a.s.}} d(P_{\boldsymbol{\theta}}, P_o)$, which is precisely the form of a Gibbs posterior as discussed in Section S1. If $R = r_0$ a.s. for some $r_0 > 0$, then $G(r) = \mathbb{1}(r < r_0)$, and by Theorem S3.3 the asymptotic c-posterior is proportional to $\mathbb{1}(d(P_\theta, P_o) < r_0)\Pi(d\theta)$, i.e., it is zero outside the radius r_0 "neighborhood" of P_o and reverts to the prior inside.

The following corollary establishes the asymptotic form of the relative entropy c-posterior.

Corollary S3.4. *Suppose P_o has density p_o, P_θ has density p_θ for each θ, and $d_n(X_{1:n}, x_{1:n})$ is an almost-surely consistent estimator of $D(p_o\|p_{\boldsymbol{\theta}})$, i.e., $d_n(X_{1:n}, x_{1:n}) \xrightarrow{\text{a.s.}} D(p_o\|p_{\boldsymbol{\theta}})$. If $d(P_\theta, P_o) = D(p_o\|p_\theta)$ and Condition S3.2 is satisfied, then Equations S3.1 and S3.2 hold.*

We also obtain the following interesting corollary. Recall that $\hat{P}_{x_{1:n}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$.

Corollary S3.5. *Suppose $d : \mathcal{M} \times \mathcal{M} \to [0, \infty]$ has the property that $d(P_n, Q_n) \to d(P, Q)$ whenever $P_n \Rightarrow P$ and $Q_n \Rightarrow Q$. If Condition S3.2 is satisfied, then Equations S3.1 and S3.2 hold when $d_n(X_{1:n}, x_{1:n}) = d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}})$.*

Proof. Since $X_1, \ldots, X_n|\boldsymbol{\theta}$ i.i.d. $\sim P_{\boldsymbol{\theta}}$ and $x_1, \ldots, x_n$ behaves like an i.i.d. sequence from P_o, then $\hat{P}_{X_{1:n}} \overset{\text{a.s.}}{\Longrightarrow} P_{\boldsymbol{\theta}}$ and $\hat{P}_{x_{1:n}} \Longrightarrow P_o$ (Dudley, 2002, Theorem 11.4.1). Hence, $d_n(X_{1:n}, x_{1:n}) \xrightarrow{\text{a.s.}} d(P_{\boldsymbol{\theta}}, P_o)$, and Theorem S3.3 applies. □

S3.2 Small-sample behavior of the c-posterior

When n is small, the c-posterior tends to be well-approximated by the standard posterior. To study this, we consider a different asymptotic regime—namely, the limit as the

distribution of R converges to 0 in a certain sense, while holding n fixed.

We continue to assume the setup from the beginning of Section S3. Further, suppose each P_θ has a density p_θ with respect to some common measure λ on $\mathcal{X}$. Let $\mathcal{E}(x_{1:n}) = \{(x_{\sigma_1}, \ldots, x_{\sigma_n}) : \sigma \in S_n\}$ where S_n is the set of permutations of $(1, \ldots, n)$ and $x_1, \ldots, x_n \in \mathcal{X}$ are the observed data.

Theorem S3.6. *There exists $c_\alpha \in (0, \infty)$ depending on $\mathcal{X}$, λ, $x_{1:n}$, d_n, and G—but not depending on θ—such that*

$$c_\alpha \, \mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R/\alpha \,\big|\, \theta\big) \xrightarrow[\alpha\to\infty]{} \prod_{i=1}^n p_\theta(x_i),$$

if either of the following two cases hold:

1. *(Discrete case) Suppose $\mathcal{X}$ is countable and λ is counting measure. Suppose $d_n(x'_{1:n}, x_{1:n}) = 0$ if and only if $x'_{1:n} \in \mathcal{E}(x_{1:n})$. Assume $G(0) > 0$.*

2. *(Continuous case) Suppose $\mathcal{X} = \mathbb{R}^m$ for some m, and λ is Lebesgue measure on $\mathcal{X}$. Assume p_θ is continuous at each of $x_1, \ldots, x_n$. Assume $d_n(x'_{1:n}, x_{1:n}) = d_n(x'_\sigma, x_{1:n})$ for all $x'_{1:n} \in \mathcal{X}^n$, $\sigma \in S_n$ (i.e., $d_n(x'_{1:n}, x_{1:n})$ is invariant to the order of $x'_1, \ldots, x'_n$). Suppose that for any sequence $x^{(1)}_{1:n}, x^{(2)}_{1:n}, \ldots \in \mathcal{X}^n$, we have $d_n(x^{(k)}_{1:n}, x_{1:n}) \to 0$ if and only if $\min_{\sigma\in S_n} \sum_{i=1}^n \|x^{(k)}_{\sigma_i} - x_i\|^2 \to 0$ as $k \to \infty$. Assume that $G(r) > 0$ for all $r \in [0, \infty)$, and that there exists $\gamma \in (0, 1)$ such that $G(r)/G(\gamma r) \to 0$ as $r \to \infty$.*

Proof. (Discrete case) By the dominated convergence theorem,

$$\begin{aligned}\frac{1/G(0)}{|\mathcal{E}(x_{1:n})|}\,\mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R/\alpha \,\big|\, \theta\big) &= \frac{1/G(0)}{|\mathcal{E}(x_{1:n})|} \sum_{x'_{1:n}\in\mathcal{X}^n} G(\alpha d_n(x'_{1:n}, x_{1:n})) \prod_{i=1}^n p_\theta(x'_i) \\ &\xrightarrow[\alpha\to\infty]{} \frac{1/G(0)}{|\mathcal{E}(x_{1:n})|} \sum_{x'_{1:n}\in\mathcal{E}(x_{1:n})} G(0) \prod_{i=1}^n p_\theta(x'_i) = \prod_{i=1}^n p_\theta(x_i).\end{aligned}$$

(Continuous case) Let us abbreviate $x = x_{1:n}$ and $\mathcal{E} = \mathcal{E}(x_{1:n})$. For $y \in \mathcal{X}^n$, denote $B_r(y) = \{z \in \mathcal{X}^n : \sum_{i=1}^n \|y_i - z_i\|^2 < r^2\}$, i.e., the Euclidean ball of radius r in $\mathbb{R}^{mn}$. Choose $r \in (0, \infty)$ small enough that for any $y, z \in \mathcal{E}$ such that $y \neq z$, we have $B_r(y) \cap$

$B_r(z) = \varnothing$. Define $\tilde{\mathcal{X}} = \big(\mathcal{X}^n \setminus \bigcup_{y\in\mathcal{E}} B_r(y)\big) \cup B_r(x)$, and give $\tilde{\mathcal{X}}$ the Euclidean metric. Define a Borel measure $\tilde{\lambda}$ on $\tilde{\mathcal{X}}$ by $\tilde{\lambda}(A) = \lambda(A) + (|\mathcal{E}| - 1)\lambda(A \cap B_r(x))$. Let $Z_\alpha = \int_{B_r(x)} G(\alpha d_n(x', x))d\tilde{\lambda}(x')$. Then

$$\frac{1}{Z_\alpha}\mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R/\alpha \,\big|\, \theta\big) \overset{\text{(a)}}{=} \frac{1}{Z_\alpha}\int_{\mathcal{X}^n} G(\alpha d_n(x', x))\big(\textstyle\prod_{i=1}^n p_\theta(x'_i)\big)dx'$$
$$\overset{\text{(b)}}{=} \frac{1}{Z_\alpha}\int_{\tilde{\mathcal{X}}} G(\alpha d_n(x', x))\big(\textstyle\prod_{i=1}^n p_\theta(x'_i)\big)d\tilde{\lambda}(x') \xrightarrow{\text{(c)}} \prod_{i=1}^n p_\theta(x_i)$$

as $\alpha \to \infty$; (a) is by Equation 2.1; (b) is since there are $|\mathcal{E}|$ distinct permutations of $x'_1, \ldots, x'_n$ and the integrand is invariant to these permutations; (c) is by Lemma S3.7 applied to $\tilde{\mathcal{X}}$, $\tilde{\lambda}$, $f(x') = d_n(x', x)$, and $h(x') = \prod_{i=1}^n p_\theta(x'_i)$. □

Lemma S3.7. *Let $\mathcal{X}$ be a metric space, and let λ be a Borel measure on $\mathcal{X}$. Let $f : \mathcal{X} \to [0, \infty]$ be measurable. Suppose there is a point $x_0 \in \mathcal{X}$ such that for any sequence $x_1, x_2, \ldots \in \mathcal{X}$, we have $f(x_n) \to 0$ if and only if $x_n \to x_0$. Let $h : \mathcal{X} \to [0, \infty)$ such that $h \in L^1(\lambda)$ and h is continuous at x_0. Assume $0 < \lambda(B_r(x_0)) < \infty$ for all $r \in (0, \infty)$. Suppose $G(r) = \mathbb{P}(R > r)$ for some random variable R on $[0, \infty)$ such that $G(r) > 0$ for all $r \in [0, \infty)$, and suppose there exists $\gamma \in (0, 1)$ such that $G(r)/G(\gamma r) \to 0$ as $r \to \infty$. Then for any $r \in (0, \infty)$,*

$$\frac{1}{Z_\alpha(r)}\int_{\mathcal{X}} G(\alpha f(x))h(x)d\lambda(x) \longrightarrow h(x_0)$$

as $\alpha \to \infty$, where $Z_\alpha(r) = \int_{B_r(x_0)} G(\alpha f(x))d\lambda(x)$.

Here, $B_r(x_0) := \{x \in \mathcal{X} : d_{\mathcal{X}}(x, x_0) < r\}$, where $d_{\mathcal{X}}$ is the metric of $\mathcal{X}$. Note that the condition on f implies, in particular, that (a) $f(x) = 0$ if and only if $x = x_0$, (b) f is continuous at x_0, and (c) for any $r > 0$, $\inf\{f(x) : x \in B_r(x_0)^c\} > 0$.

Proof. Let us abbreviate $B_r = B_r(x_0)$. Let $\varepsilon > 0$. Using the continuity of h at x_0, choose $\delta \in (0, r)$ such that for all $x \in B_\delta$, $h(x_0) - \varepsilon \le h(x) \le h(x_0) + \varepsilon$. Let $\alpha > 0$. Then

$$\frac{1}{Z_\alpha(r)}\int_{\mathcal{X}} G(\alpha f(x))h(x)d\lambda = \frac{1}{Z_\alpha(r)}\int_{B_\delta^c} G(\alpha f(x))h(x)d\lambda + \frac{Z_\alpha(\delta)}{Z_\alpha(r)}\frac{1}{Z_\alpha(\delta)}\int_{B_\delta} G(\alpha f(x))h(x)d\lambda.$$

By our choice of δ,

$$h(x_0) - \varepsilon \leq \frac{1}{Z_\alpha(\delta)} \int_{B_\delta} G(\alpha f(x)) h(x) d\lambda \leq h(x_0) + \varepsilon.$$

So, if we can show that $\frac{1}{Z_\alpha(r)} \int_{B_\delta^c} G(\alpha f(x)) h(x) d\lambda \to 0$ and $Z_\alpha(\delta)/Z_\alpha(r) \to 1$ as $\alpha \to \infty$, then the result will follow since $\varepsilon > 0$ is arbitrary. Let $\beta = \min\{1, \inf\{f(x) : x \in B_\delta^c\}\}$, and note that $0 < \beta < \infty$. By the continuity of f at x_0, choose $\rho \in (0, \delta)$ such that $f(x) < \beta\gamma$ for all $x \in B_\rho$. Then

$$0 \leq \frac{1}{Z_\alpha(r)} \int_{B_\delta^c} G(\alpha f(x)) h(x) d\lambda \leq \frac{G(\alpha\beta) \int_{B_\delta^c} h(x) d\lambda}{\int_{B_\rho} G(\alpha f(x)) d\lambda} \leq \frac{G(\alpha\beta) \int_{\mathcal{X}} h(x) d\lambda}{G(\alpha\beta\gamma)\lambda(B_\rho)} \longrightarrow 0$$

as $\alpha \to \infty$. Similarly,

$$\frac{Z_\alpha(r)}{Z_\alpha(\delta)} = 1 + \frac{\int_{B_r \setminus B_\delta} G(\alpha f(x)) d\lambda}{\int_{B_\delta} G(\alpha f(x)) d\lambda} \longrightarrow 1.$$

□

S3.3 Robustness of the c-posterior

The definition of robustness, roughly speaking, is that small changes to the distribution of the data result in small changes to the resulting inferences. This can be formalized by requiring that the outcome of an inference procedure be continuous as a function of P_o, asymptotically, with respect to some topology (the weak topology being a standard choice) (Huber, 2004). The lack of robustness of the standard posterior can be seen as a lack of continuity with respect to P_o, asymptotically (see Section S4).

We show in the following theorem that the asymptotic c-posterior inherits the continuity properties of whatever discrepancy $d(\cdot, \cdot)$ is used to define it. Consequently, the c-posterior is robust to perturbations to P_o that are small with respect to $d(\cdot, \cdot)$. In the terminology of Section 2, if the observed data distribution P_o is close to the idealized distribution P_{θ_I}, then the c-posterior will be close to what it would be if $P_o = P_{\theta_I}$.

To interpret the theorem, recall that on any metric space, a function $f(x)$ is continuous if and only if $x_m \to x$ implies $f(x_m) \to f(x)$. Thus, to show continuity as a function

of P_o (in some topology), one must show that if $P_m \to P_o$, then the resulting sequence of asymptotic c-posteriors converges as well. In fact, if $d(\cdot, \cdot)$ is continuous (in this same topology), then the theorem shows a bit more than that, since then $P_m \to P_o$ implies $d(P_\theta, P_m) \to d(P_\theta, P_o)$.

Theorem S3.8. *If $P_1, P_2, \ldots \in \mathcal{M}$ such that $d(P_\theta, P_m) \to d(P_\theta, P_o)$ as $m \to \infty$ for Π-almost all $\theta \in \Theta$, and Condition S3.2 is satisfied, then for any $h \in L^1(\Pi)$,*

$$\mathbb{E}\big(h(\boldsymbol{\theta}) \mid d(P_{\boldsymbol{\theta}}, P_m) < R\big) \longrightarrow \mathbb{E}\big(h(\boldsymbol{\theta}) \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big)$$

as $m \to \infty$, and in particular, $\Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_m) < R\big) \Longrightarrow \Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big)$.

Proof. Apply Lemma S3.1 with $U = d(P_{\boldsymbol{\theta}}, P_o)$, $U_m = d(P_{\boldsymbol{\theta}}, P_m)$, $V = R$, and $W = h(\boldsymbol{\theta})$. □

Theorem S3.8 implies that the c-posterior is robust in the context of model selection/inference, since if $h(\theta) = \mathbb{1}(\theta \in \Theta_k)$ where Θ_k represents model k (see Section S4), then $\Pi\big(\Theta_k \mid d(P_{\boldsymbol{\theta}}, P_m) < R\big) \longrightarrow \Pi\big(\Theta_k \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big)$ as $m \to \infty$, under the assumptions of the theorem.

The statement of the theorem concerns the asymptotic c-posterior, rather than the finite-sample c-posterior, because the characterization of robustness in terms of continuity only makes sense asymptotically. A similar result can be proved in the finite-sample case, but this would be uninteresting since usually the standard posterior is also continuous (but perhaps highly sensitive) with respect to a finite sample. What would be more interesting in the finite-sample case would be to quantify or bound the change in posterior expectations of interest, as a function of the size of the perturbation.

S4 Lack of robustness of the standard posterior

Standard model selection methods do not address the model misspecification/perturbation problem, because they choose the model that is nearest in Kullback–Leibler divergence to

the observed data distribution, when n is sufficiently large. We focus here on Bayesian model averaging, but similar arguments apply to AIC and BIC, for example.

When n is large, the standard posterior can be strongly affected by small changes to the observed data distribution P_o, particularly when performing model selection/inference (see Section S4.1), while c-posteriors are robust to small changes in P_o (as shown in Section S3.3). To see roughly why the standard posterior is not robust, note that if P_o and P_θ have densities p_o and p_θ, respectively, and the prior Π has density π, then

$$\pi(\theta \mid X_{1:n} = x_{1:n}) \propto \exp\Big(\sum_{i=1}^{n} \log p_\theta(x_i)\Big)\pi(\theta) \doteq \exp\big(n \textstyle\int p_o \log p_\theta\big)\pi(\theta)$$
$$\propto \exp(-nD(p_o\|p_\theta))\pi(\theta),$$

where $\doteq$ denotes agreement to first order in the exponent (in other words, $a_n \doteq b_n$ if $(1/n)\log(a_n/b_n) \to 0$). Due to the n in the exponent, even a slight change to p_o can dramatically change the posterior. On the other hand, by comparison, the relative entropy c-posterior with $R \sim \mathrm{Exp}(\alpha)$ is asymptotically proportional to $\exp(-\alpha D(p_o\|p_\theta))\pi(\theta)$, and consequently, it remains stable in the limit as $n \to \infty$ (Section S3).

S4.1 Model selection sensitivity

The standard posterior is particularly susceptible to robustness issues when applied to model selection/inference. Suppose that for each k in some countable index set, we have a model $\mathcal{M}_k = \{P_\theta : \theta \in \Theta_k\}$, where Θ_k is a t_k-dimensional Euclidean space. Let $\pi(k)$ be a prior on the model index k, and for each k, let π_k be a probability density with respect to Lebesgue measure on Θ_k; this induces a prior Π on the disjoint union $\Theta = \bigcup_k \Theta_k$.

It is well-known that, under mild regularity conditions, the marginal likelihood $p(x_{1:n}|k) = \int_{\Theta_k} p(x_{1:n}|\theta)\pi_k(\theta)d\theta$ has the asymptotic representation

$$p(x_{1:n}|k) \sim \frac{p(x_{1:n}|\theta_k^n)\pi_k(\theta_k^*)}{|\det H(\theta_k^*; p_o)|^{1/2}}\Big(\frac{2\pi}{n}\Big)^{t_k/2},$$

as $n \to \infty$, where $\theta_k^n = \mathrm{argmax}_{\theta\in\Theta_k} p(x_{1:n}|\theta)$ is the maximum likelihood estimator for model k, $\theta_k^* = \mathrm{argmin}_{\theta\in\Theta_k} D(p_o\|p_\theta)$ is the minimal Kullback–Leibler (KL) point within

model k, and $H(\theta; p_o) = -\int p_o\big(\nabla^2_\theta \log p_\theta\big)$. Here, $a_n \sim b_n$ means $a_n/b_n \to 1$. Letting $f_n(k) = -\frac{1}{n}\log p(x_{1:n}|\theta^n_k)$, this implies that

$$p(x_{1:n}|k) \sim c_k e^{-nf_n(k)} n^{-t_k/2} \tag{S4.1}$$

for a constant c_k not depending on n or $x_{1:n}$. Typically, $f_n(k) \to f(k) := D(p_o \| p_{\theta^*_k}) - \int p_o \log p_o$. Note that $f(k') < f(k)$ if and only if model k' is closer to p_o than model k in terms of minimal KL divergence; also, note that the marginal likelihood automatically penalizes more complex models via the $n^{-t_k/2}$ factor.

Given such an asymptotic representation, it is easy to see that for any k, if there exists k' such that $f(k') < f(k)$, then $\pi(k|x_{1:n}) \to 0$ as $n \to \infty$. Consequently, even the slightest change to p_o can result in major shifts in the posterior on k, when n is large. For instance, it often happens that the models are nested, e.g., $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \cdots$ and $t_1 < t_2 < \cdots$. This is the case, for example, when $\mathcal{M}_k$ consists of k-component mixtures, or kth-order autoregressive models; variable selection is slightly more complicated but ultimately similar. If the collection of models is correctly specified with respect to p_o, then there is some minimal k' such that $D(p_o \| p_{\theta^*_{k'}}) = 0$, and thus $\pi(k|x_{1:n}) \to 0$ for all $k < k'$ (and typically, the posterior on k will concentrate at this k'). However, even the slightest perturbation to p_o will usually result in either (a) an increase in this minimal k', or (b) a situation where $\inf_k D(p_o \| p_{\theta^*_k})$ is not attained at any k, causing the posterior on k to diverge, in the sense that $\pi(k|x_{1:n}) \to 0$ for all k. Hence, model selection/inference with the standard posterior is not robust.

S5 Power posterior approximation

In this section, we provide further explanation of the power posterior approximation to the relative entropy c-posterior. As shown in Section S3.1, if $d_n(X_{1:n}, x_{1:n})$ is a consistent estimator of $D(p_o \| p_\theta)$, and $R \sim \text{Exp}(\alpha)$, then asymptotically as $n \to \infty$, the c-posterior

based on $d_n(X_{1:n}, x_{1:n})$ is proportional to

$$\begin{aligned}\exp(-\alpha D(p_o \| p_\theta))\,\pi(\theta) &\propto \exp(\alpha \textstyle\int p_o \log p_\theta)\,\pi(\theta) \\ &\approx \exp\Big(\alpha \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i)\Big)\,\pi(\theta) = \pi(\theta)\prod_{i=1}^{n} p_\theta(x_i)^{\alpha/n}\end{aligned}$$

under mild regularity conditions. Thus, the power posterior approximation in Equation 2.2 is good when $n \gg \alpha$, since then $\zeta_n \approx \alpha/n$.

Meanwhile, by Theorem S3.6, when $\alpha \gg n$ the c-posterior is well-approximated by the standard posterior, under regularity conditions. Thus, since $\zeta_n \approx 1$ when $\alpha \gg n$, the power posterior approximation is also good when n is much smaller than α. This makes intuitive sense since the distribution of R is strongly concentrated near 0 when $\alpha \gg n$, and thus, for an appropriate choice of d_n, conditioning on $d_n(X_{1:n}, x_{1:n}) < R$ is roughly the same as conditioning on the event that $X_{1:n}$ and $x_{1:n}$ have the same empirical distribution.

What about the intermediate regime where n and α are comparable in magnitude? The point of choosing $\zeta_n = \alpha/(\alpha + n)$ is that it smoothly transitions through this intermediate regime; for this reason we refer to it as the power interpolation formula. This particular formula for ζ_n is obtained by analyzing the special case where the sample space $\mathcal{X}$ has finitely many elements. When $|\mathcal{X}| < \infty$, a natural choice of $d_n(X_{1:n}, x_{1:n})$ is simply $D(\hat{p}_{x_{1:n}} \| \hat{p}_{X_{1:n}})$, that is, the relative entropy of the empirical densities. If, further, $R \sim \text{Exp}(\alpha)$, then by an approximation detailed in Section S5.1,

$$\begin{aligned}\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) &\propto \mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \theta\big)\,\pi(\theta) \\ &= \mathbb{E}\big(\exp(-\alpha D(\hat{p}_{x_{1:n}} \| \hat{p}_{X_{1:n}})) \mid \theta\big)\,\pi(\theta) \\ &\approx (n\zeta_n/\alpha)^{\frac{|\mathcal{X}|-1}{2}} \exp(-n\zeta_n D(\hat{p}_{x_{1:n}} \| p_\theta))\,\pi(\theta) \qquad \text{(S5.1)} \\ &\propto \pi(\theta)\prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n}\end{aligned}$$

where $\propto$ indicates proportionality with respect to θ, and $\zeta_n = \alpha/(\alpha + n)$.

S5.1 Justification of Equation S5.1

Let $\Delta_k = \{p \in \mathbb{R}^k : \sum_i p_i = 1,\ p_i > 0\ \forall i\}$. Let $s \in \Delta_k$. We argue that if $X_1, \ldots, X_n$ i.i.d. $\sim s$ and $\hat{\boldsymbol{s}}_j = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i = j)$ for $j = 1, \ldots, k$, then for $p \in \Delta_k$ near s,

$$\mathbb{E}\exp(-\alpha D(p\|\hat{\boldsymbol{s}})) \approx (n\zeta_n/\alpha)^{\frac{k-1}{2}} \exp(-n\zeta_n D(p\|s)), \tag{S5.2}$$

where $\zeta_n = (1/n)/(1/n + 1/\alpha)$. We use bold here to denote random variables. For $x \in \mathbb{R}^d$, define $C(x) \in \mathbb{R}^{d\times d}$ such that $C(x)_{ij} = x_i\mathbb{1}(i = j) - x_i x_j$, and denote $x' = (x_1, \ldots, x_{d-1})$. First, for $q \in \Delta_k$ near p,

$$D(p\|q) \approx \tfrac{1}{2}\chi^2(p, q) = \tfrac{1}{2}(p' - q')^{\mathsf{T}} C(q')^{-1}(p' - q') \tag{S5.3}$$

by Propositions S5.1 and S5.2 below. By the central limit theorem, $\hat{\boldsymbol{s}}$ is approximately $\mathcal{N}(s, C(s)/n)$ distributed. Therefore, letting $\boldsymbol{q} \sim \mathcal{N}(s, C(s)/n)$ and $C = C(s')$,

$$\begin{aligned}
\mathbb{E}\exp(-\alpha D(p\|\hat{\boldsymbol{s}})) &\approx \mathbb{E}\exp(-\alpha D(p\|\boldsymbol{q}))\mathbb{1}(\boldsymbol{q} \in \Delta_k) \\
&\overset{\text{(a)}}{\approx} \mathbb{E}\exp\Big(-\frac{\alpha}{2}(p' - \boldsymbol{q}')^{\mathsf{T}} C^{-1}(p' - \boldsymbol{q}')\Big) \\
&= (2\pi)^{\frac{k-1}{2}}|C/\alpha|^{1/2}\int \mathcal{N}(p'|q', C/\alpha)\mathcal{N}(q'|s', C/n)dq' \\
&\overset{\text{(b)}}{=} (2\pi)^{\frac{k-1}{2}}|C/\alpha|^{1/2}\mathcal{N}(p'|s', (1/\alpha + 1/n)C) \\
&= \Big(\frac{1/\alpha}{1/\alpha + 1/n}\Big)^{\frac{k-1}{2}} \exp\big(-\tfrac{1}{2}(1/\alpha + 1/n)^{-1}(p' - s')^{\mathsf{T}} C^{-1}(p' - s')\big) \\
&\overset{\text{(c)}}{\approx} (n\zeta_n/\alpha)^{\frac{k-1}{2}} \exp(-n\zeta_n D(p\|s)),
\end{aligned}$$

where (a) is by Equation S5.3 along with the approximation $C(\boldsymbol{q}') \approx C(s')$, (b) uses the convolution formula for independent normals, and (c) is again by Equation S5.3. This yields Equation S5.2.

It is well-known that chi-squared distance is a second-order Taylor approximation to relative entropy (Cover and Thomas, 2006, Lemma 17.3.3); for completeness, we include the proof.

Proposition S5.1. *For $p, q \in \Delta_k$, $D(p\|q) = \frac{1}{2}\chi^2(p, q) + o(\|p - q\|^2)$ as $p \to q$, where $D(p\|q) = \sum_i p_i \log(p_i/q_i)$ and $\chi^2(p, q) = \sum_i (p_i - q_i)^2/q_i$.*

Proof. Fix $b > 0$, and define $f(a) = a \log(a/b)$ for $a > 0$. Then by Taylor's theorem,

$$\begin{aligned} f(a) &= f(b) + f'(b)(a-b) + \tfrac{1}{2} f''(b)(a-b)^2 + o(|a-b|^2) \\ &= (a-b) + \frac{1}{2}\frac{(a-b)^2}{b} + o(|a-b|^2) \end{aligned}$$

as $a \to b$. It follows that

$$\sum_{i=1}^{k} p_i \log \frac{p_i}{q_i} = \sum_i (p_i - q_i) + \frac{1}{2} \sum_i \frac{(p_i - q_i)^2}{q_i} + o(\|p-q\|^2) = \tfrac{1}{2}\chi^2(p,q) + o(\|p-q\|^2)$$

as $p \to q$. □

The following result expresses the chi-squared distance $\chi^2(p,q)$ in terms of the $(k-1)$-dimensional Mahalanobis distance for Z' when $Z \sim \text{Multinomial}(1, q)$. For interpretation, note that C below equals $\text{Cov}(Z')$ when $Z \sim \text{Multinomial}(1, q)$.

Proposition S5.2. *For any $p, q \in \Delta_k$, $\chi^2(p,q) = (p' - q')^\intercal C^{-1}(p' - q')$ where $C \in \mathbb{R}^{(k-1)\times(k-1)}$ such that $C_{ij} = q_i \mathbb{1}(i = j) - q_i q_j$.*

Proof. By the Sherman–Morrison formula for rank-one updates, $C^{-1} = (\text{diag}(q') - q'q'^\intercal)^{-1} = \text{diag}(q')^{-1} + (1/q_k)\mathbf{1}\mathbf{1}^\intercal$ where $\mathbf{1} = (1, \ldots, 1)^\intercal$, hence

$$(p' - q')^\intercal C^{-1}(p' - q') = \sum_{i=1}^{k-1} \frac{(p_i - q_i)^2}{q_i} + \frac{\left(\sum_{i=1}^{k-1}(p_i - q_i)\right)^2}{q_k}$$

and $\sum_{i=1}^{k-1}(p_i - q_i) = (1 - p_k) - (1 - q_k) = q_k - p_k$. □

S6 Extensions

S6.1 Time-series c-posterior based on relative entropy rate

Suppose the sequence of observed data $(x_1, \ldots, x_n)$ is a partial sample from a stationary and ergodic process with distribution P_o, and suppose the model $\{P_\theta : \theta \in \Theta\}$ consists of stationary finite-order Markov processes. Assume that for some sigma-finite measure μ on $\mathcal{X}$, for all $n \in \{1, 2, \ldots\}$ and all $\theta \in \Theta$, the finite-dimensional distributions have

densities $p_o(x_1, \ldots, x_n)$ and $p_\theta(x_1, \ldots, x_n)$ with respect to the product measure μ^n, and assume $\mathbb{E}_{P_o}|\log p_o(X_{1:n})| < \infty$ and $\mathbb{E}_{P_o}|\log p_\theta(X_{1:n})| < \infty$.

A natural way of assessing the discrepancy between the processes P_o and P_θ is by the relative entropy rate (Gray, 1990),

$$\mathcal{D}(P_o\|P_\theta) = \lim_{n\to\infty} \frac{1}{n} D\big(p_o(x_{1:n})\|p_\theta(x_{1:n})\big).$$

Suppose $d_n(X_{1:n}, x_{1:n})$ is an a.s.-consistent estimator of $\mathcal{D}(P_o\|P_\theta)$ when $(X_1, X_2, \ldots) \sim P_\theta$ and $(x_1, x_2, \ldots) \sim P_o$, and consider the c-posterior $\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$, with $R \sim \text{Exp}(\alpha)$. Then by Lemma S3.1, the asymptotic c-posterior is

$$\Pi(d\theta \mid \mathcal{D}(P_o\|P_\theta) < R) \propto \exp(-\alpha\mathcal{D}(P_o\|P_\theta))\Pi(d\theta).$$

If P_θ is kth-order Markov, then

$$\mathcal{D}(P_o\|P_\theta) = -\mathcal{H}(P_o) - \mathbb{E}_{P_o} \log p_\theta(X_{k+1}|X_1, \ldots, X_k)$$

where $\mathcal{H}(P_o)$ is the entropy rate of P_o, which we assume is finite (Gray, 1990, Lemma 2.4.3). Further, when $(x_1, x_2, \ldots) \sim P_o$,

$$\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i|x_1, \ldots, x_{i-1}) \xrightarrow[n\to\infty]{} \mathbb{E}_{P_o} \log p_\theta(X_{k+1}|X_1, \ldots, X_k)$$

with probability 1, by the ergodic theorem (Breiman, 1968, 6.28). This leads to the approximation

$$\begin{aligned}\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) &\approx\!\!\propto \exp\Big(-n\zeta_n\Big[-\mathcal{H}(P_o) - \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x_i|x_1, \ldots, x_{i-1})\Big]\Big)\Pi(d\theta) \\ &\propto \Pi(d\theta)\prod_{i=1}^{n} p_\theta(x_i|x_1, \ldots, x_{i-1})^{\zeta_n},\end{aligned}$$

using the power interpolation formula $\zeta_n = \alpha/(\alpha + n)$ to scale appropriately for small n. Thus, as in the i.i.d. case, the end result is an approximation obtained by simply raising the likelihood to the power ζ_n. In Section 6, we apply this to perform robust inference for the order of an autoregressive model.

S6.2 Regression c-posterior based on conditional relative entropy

In regression, one observes covariates/predictors $x_1, \ldots, x_n$ associated with target values $y_1, \ldots, y_n$, and models the conditional distribution of y given x. As in the i.i.d. setting, in order to construct a c-posterior allowing for perturbations/misspecification, let us assume that $Y_i|x_i$ is drawn from the model $p_\theta(y|x)$ for $i = 1, \ldots, n$, and that the observed values $y_{1:n}$ are a slightly perturbed version of $Y_{1:n}$, in the sense that $d_n(Y_{1:n}, y_{1:n}|x_{1:n}) < R$ for some measure of discrepancy $d_n(\cdot, \cdot|\cdot)$. Suppose that, in fact, the observed data $(x_1, y_1), \ldots, (x_n, y_n)$ behave like i.i.d. samples from some $p_o(x, y)$. For notational clarity, let us assume that these densities on x and y are with respect to measures that we will denote by dx and dy, respectively.

A natural choice of discrepancy between the conditional distributions $p_o(y|x)$ and $p_\theta(y|x)$ is the conditional relative entropy,

$$D_\theta := \int p_o(x, y) \log \frac{p_o(y|x)}{p_\theta(y|x)} \, dx \, dy,$$

and in turn, an a.s.-consistent estimator of this quantity is a sensible choice for $d_n(\cdot, \cdot|\cdot)$. Then, by Lemma S3.1, the resulting c-posterior converges to a nice asymptotic form:

$$\begin{aligned}\Pi\big(d\theta \mid d_n(Y_{1:n}, y_{1:n}|x_{1:n}) < R\big) \Longrightarrow \Pi\left(d\theta \mid D_\theta < R\right) &\propto \exp(-\alpha D_\theta)\Pi(d\theta) \\ &\propto \exp\Big(\alpha \int p_o(x, y) \log p_\theta(y|x) \, dx \, dy\Big)\Pi(d\theta)\end{aligned}$$

if we take $R \sim \mathrm{Exp}(\alpha)$ as usual. To obtain an approximation that is applicable for smaller n as well, we apply the same power interpolation formula as before, replacing α by $n\zeta_n$. Along with an empirical approximation to the integral, this suggests using

$$\begin{aligned}\Pi(d\theta \mid d_n(Y_{1:n}, y_{1:n}|x_{1:n}) < R) &\approx\!\!\propto \exp\big(\zeta_n \textstyle\sum_i \log p_\theta(y_i|x_i)\big)\Pi(d\theta) \\ &= \Pi(d\theta) \prod_{i=1}^n p_\theta(y_i|x_i)^{\zeta_n}.\end{aligned}$$

Consequently, once again, we arrive at a power posterior approximation to the c-posterior, allowing us to bypass the computation of $d_n(\cdot, \cdot|\cdot)$. In Section 7, we apply this to perform robust variable selection in linear regression.

S7 Additional details on the examples

S7.1 Bernoulli example details

S7.1.1 Computation of the exact c-posterior

Letting $Z = \mathbb{1}(D(\hat{p}_x||\hat{p}_X) < R)$, by Bayes' theorem we have that for $h \in \{\mathrm{H}_0, \mathrm{H}_1\}$,

$$\begin{aligned}\Pi\big(h|Z=1\big) &\underset{h}{\propto} \mathbb{P}(Z=1|h)\Pi(h) \overset{\text{(a)}}{\underset{h}{\propto}} \mathbb{P}(Z=1|h) \\ &\overset{\text{(b)}}{=} \mathbb{E}\big(\mathbb{P}(Z=1|X_{1:n},h)\,\big|\,h\big) \overset{\text{(c)}}{=} \mathbb{E}\big(\exp(-\alpha D(\hat{p}_x||\hat{p}_X))\,\big|\,h\big)\end{aligned}$$

where (a) is since $\Pi(h) = 1/2$, (b) is by the law of iterated expectations, and (c) by the fact that $\mathbb{P}(R > r) = \exp(-\alpha r)$. This is easily computed exactly, since, letting $S = \sum_{i=1}^n X_i = n\hat{p}_X(1)$, we have $S|\mathrm{H}_0 \sim \text{Binomial}(n, 1/2)$ and $S|\mathrm{H}_1 \sim \text{BetaBinomial}(n, 1, 1) = \text{Uniform}\{0, 1, \ldots, n\}$.

S7.1.2 Formulas for the power posterior and the calibration curve

The power posterior on (h, θ), for $h \in \{\mathrm{H}_0, \mathrm{H}_1\}$ and $\theta \in (0, 1)$, is

$$\pi_\alpha(h, \theta|x_{1:n}) = c\,\pi(h,\theta)\prod_{i=1}^n p_\theta(x_i)^\zeta = c\,\pi(h)\pi(\theta|h)\,\theta^{s\zeta}(1-\theta)^{(n-s)\zeta}$$

where $p_\theta(x) = \theta^x(1-\theta)^{1-x}\mathbb{1}(x \in \{0,1\})$, $\zeta = \alpha/(\alpha+n)$, and $s = \sum_{i=1}^n x_i$. If $h = \mathrm{H}_0$, then $\theta = 1/2$ with probability 1, so integrating over θ yields

$$\Pi_\alpha(\mathrm{H}_0|x_{1:n}) = c\,\Pi(\mathrm{H}_0)(1/2)^{s\zeta}(1-1/2)^{(n-s)\zeta} = c\,(1/2)^{n\zeta+1}.$$

Meanwhile, if $h = \mathrm{H}_1$, then $\pi(\theta|h) = \text{Uniform}(\theta|0,1)$, so

$$\Pi_\alpha(\mathrm{H}_1|x_{1:n}) = c\,\Pi(\mathrm{H}_1)\int_0^1 \theta^{s\zeta}(1-\theta)^{(n-s)\zeta}d\theta = c\,\tfrac{1}{2}B(s\zeta+1,(n-s)\zeta+1).$$

Therefore, $\Pi_\alpha(\mathrm{H}_0|x_{1:n}) = 1/\big(1+2^{n\zeta}B(s\zeta+1,(n-s)\zeta+1)\big)$, which establishes Equation 3.2.

For the α calibration curve, we need the expected log likelihood under the power posterior, and it turns out that this can also be computed analytically. Note that $\pi_\alpha(\theta|\mathrm{H}_1, x_{1:n}) \propto$

$\pi(\theta|\mathrm{H}_1)\,\theta^{s\zeta}(1-\theta)^{(n-s)\zeta} \propto \mathrm{Beta}(\theta|A,B)$ where $A = s\zeta+1$ and $B = (n-s)\zeta+1$. Therefore,

$$\begin{aligned}\int \big(\log p(x_{1:n}|\theta)\big)\Pi_\alpha(d\theta|x_{1:n}) &= \mathrm{E}_\alpha\big(\textstyle\sum_i \log p_\theta(x_i) \,\big|\, x_{1:n}\big)\\ &= \mathrm{E}_\alpha\Big(\mathrm{E}_\alpha\big(s\log\theta + (n-s)\log(1-\theta) \mid H, x_{1:n}\big)\,\Big|\, x_{1:n}\Big)\\ &= (n\log\tfrac{1}{2})\Pi_\alpha(\mathrm{H}_0|x_{1:n}) + \big(s\,G(A,B) + (n-s)G(B,A)\big)\Pi_\alpha(\mathrm{H}_1|x_{1:n}),\end{aligned}$$

where $G(a,b) = \int(\log\theta)\mathrm{Beta}(\theta|a,b)d\theta = \psi(a) - \psi(a+b)$, with $\psi(\cdot)$ denoting the digamma function.

S7.2 Mixture model details

S7.2.1 Motivation behind the conditional coarsening algorithm

Here, we explain why Algorithm 5.1 behaves similarly to (but not exactly the same as) sampling from the power posterior. Suppose the power $\zeta \in (0,1)$ is such that $n\zeta$ is an integer, say, $m = n\zeta$. Recall that using the power posterior can be thought of as reducing the sample size to $m = n\zeta$, in terms of the concentration of the posterior. In fact, the power posterior is proportional to the geometric mean of all the posteriors based on subsets of size m, that is, letting $S = \big\{A \subseteq \{1,\ldots,n\} : |A| = m\big\}$, we have

$$\begin{aligned}\prod_{A\in S}\pi(\theta|x_A)^{1/|S|} \propto \prod_{A\in S}\big(p(x_A|\theta)\pi(\theta)\big)^{1/|S|} &= \pi(\theta)\prod_{A\in S}\prod_{j\in A}p(x_j|\theta)^{1/|S|} \qquad \text{(S7.1)}\\ &= \pi(\theta)\prod_{j=1}^{n}p(x_j|\theta)^{N_j/|S|} = \pi(\theta)\prod_{j=1}^{n}p(x_j|\theta)^{\zeta}\end{aligned}$$

where $N_j = \#\{A \in S : j \in A\}$; the last step holds since $|S| = \binom{n}{m}$ and $N_j = \binom{n-1}{m-1}$, hence $N_j/|S| = m/n = \zeta$ for all j.

Now, if A is uniformly drawn from S, then $\pi(\theta|x_A)$ provides a noisy approximation to the power posterior. So one could simply choose a random subset A of size m, and apply the standard Gibbs sampling algorithm as though x_A were all of the data, i.e., apply Algorithm 5.1 with input data $x_A = (x_j : j \in A)$ and with ζ_n set to 1. However, this would be more noisy than the power posterior, since the power posterior uses all of the available data, rather than just a fraction of it.

To obtain a better approximation, we propose to aggregate over all $A \in S$ when performing the parameter updates in the Gibbs sampling algorithm. Specifically, update the assignment variables $z_1, \ldots, z_n$ in the usual way (by sampling $z_j \sim \text{Categorical}(\tilde{w})$ where $\tilde{w}_i \propto w_i f_{\varphi_i}(x_j)$), but then update the mixture parameters (w, φ) using the geometric mean of all the full conditional distributions based on subsets $A \in S$, i.e., update w and φ using the distribution defined by $q(w, \varphi) \propto \prod_{A \in S} \pi(w, \varphi | x_A, z_A)^{1/|S|}$. By the same logic as in Equation S7.1, this is equivalent to $q(w, \varphi) \propto \pi(w, \varphi) \prod_{j=1}^{n} \left(w_{z_j} f_{\varphi_{z_j}}(x_j)\right)^{\zeta}$. Thus, we arrive at the w and φ update steps in Algorithm 5.1 by sampling from $q(w, \varphi)$, or more precisely, by making moves that preserve $q(w, \varphi)$.

S7.2.2 Importance sampling algorithm

To evaluate how well the conditional coarsening algorithm approximates the power posterior, we consider the following importance sampling (IS) algorithm. Uniformly at random, choose a subset $A \subseteq \{1, \ldots, n\}$ of size m, where m is the nearest integer to $n\zeta_n$. Run Algorithm 5.1 with input data $x_A = (x_j : j \in A)$ and with ζ_n set to 1, in order to draw MCMC samples from $\pi(w, \varphi | x_A)$, i.e., the standard posterior for this subset of points. Then, reweight the MCMC samples using importance sampling weights with the power posterior $\pi_\alpha(w, \varphi | x_{1:n})$ as the target distribution and $\pi(w, \varphi | x_A)$ as the proposal distribution. Note that the normalizing constants do not need to be evaluated for importance sampling.

This IS algorithm is guaranteed to converge to the power posterior, however, it does not scale well since $\pi(w, \varphi | x_A)$ does not always remain sufficiently close to the power posterior as the sample size and dimensionality grow, causing the effective sample size to degrade. Nonetheless, it allows us to compare results with the conditional coarsening algorithm in limited cases where the effective sample size is acceptable. Figure S1 shows the IS algorithm results on the perturbed univariate Gaussian mixture examples from Section 5.1. To facilitate comparison, we use the same values of α as in Section 5.1, namely, $\alpha = 800$ for the $k_0 = 2$ example and $\alpha = 2000$ for the $k_0 = 4$ example. The IS algorithm results are quite similar to the conditional coarsening algorithm results (compare with Figure 3),

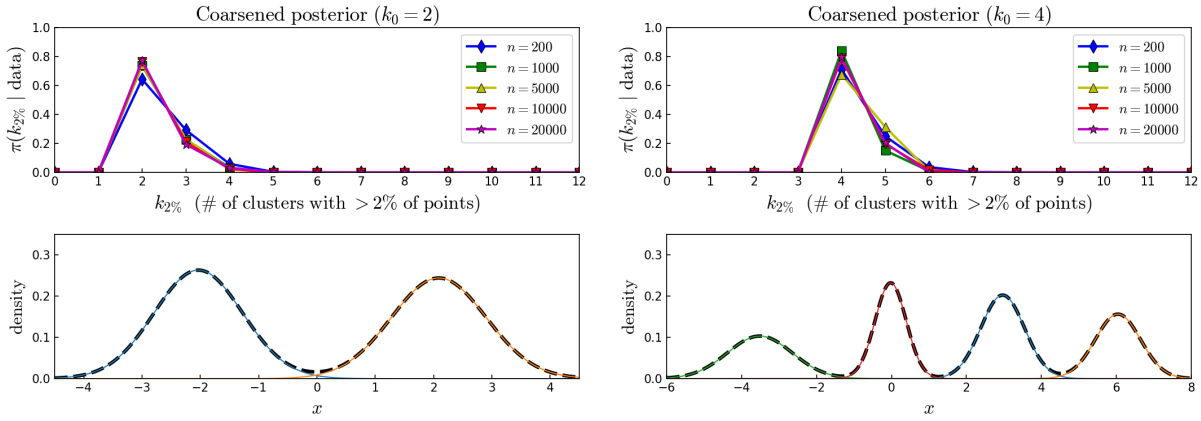


Figure S1: Top: Using the IS algorithm, the c-posterior favors the true number of nonnegligible clusters in the both examples. Bottom: Mixture density (dotted black line) and components (solid colors) for prototypical samples from the c-posterior, in each example, when $n = 20000$.

except that the conditional coarsening results tend to be slightly more concentrated.

S7.3 Normal example details

S7.3.1 Algorithm derivation

We derive Algorithm 8.1. The target distribution is $q(\theta, \alpha) = \pi\big(\theta, \alpha \mid d_n(X_{1:n}, x_{1:n}) < R\big)$, where $\theta = (\mu, \sigma^2)$. Define $h_\theta(z_{1:n}) = z_{1:n}\sigma + \mu$, so that $X_{1:n} = h_\theta(Z_{1:n})$ given θ, where $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$ independently. As in Equation 2.1,

$$\begin{aligned} q(\theta, \alpha) &\propto \pi(\theta)\pi(\alpha)\mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \theta, \alpha\big) \\ &= \pi(\theta)\pi(\alpha)\mathbb{P}\big(d_n(h_\theta(Z_{1:n}), x_{1:n}) < R \mid \theta, \alpha\big) \\ &= \pi(\theta)\pi(\alpha)\int w_{\theta,\alpha}(z_{1:n})\pi(z_{1:n})dz_{1:n} \end{aligned}$$

where $w_{\theta,\alpha}(z_{1:n}) = \exp\big(-\alpha d_n(h_\theta(z_{1:n}), x_{1:n})\big)$ and $\pi(z_{1:n}) = \prod_{i=1}^n \mathcal{N}(z_i|0, 1)$. Augment the space by introducing auxiliary variables $z_{1:n}$ with conditional distribution $q(z_{1:n}|\theta, \alpha) \propto w_{\theta,\alpha}(z_{1:n})\pi(z_{1:n})$. Then the joint distribution is $q(z_{1:n}, \theta, \alpha) \propto w_{\theta,\alpha}(z_{1:n})\pi(z_{1:n})\pi(\theta)\pi(\alpha)$. The algorithm uses Gibbs sampling for α and Metropolis-within-Gibbs for θ and $z_{1:n}$:

1. The full conditional of α is

$$q(\alpha|z_{1:n},\theta) \propto w_{\theta,\alpha}(z_{1:n})\pi(\alpha) \propto \text{Gamma}\big(\alpha \mid u,\, v + d_n(h_\theta(z_{1:n}), x_{1:n})\big).$$

2. A Metropolis move preserving $q(\theta|z_{1:n},\alpha)$ is made by proposing $\theta' \sim \mathcal{N}(\theta, \varepsilon_\theta^2 I)$ and accepting with probability $\min\left\{1,\, \dfrac{w_{\theta',\alpha}(z_{1:n})\pi(\theta')}{w_{\theta,\alpha}(z_{1:n})\pi(\theta)}\right\}$.

3. A Metropolis move preserving $q(z_{1:n}|\theta,\alpha)$ is made by proposing $z'_{1:n} \sim \mathcal{N}(z_{1:n}, \varepsilon_z^2 I)$ and accepting with probability $\min\left\{1,\, \dfrac{w_{\theta,\alpha}(z'_{1:n})\pi(z'_{1:n})}{w_{\theta,\alpha}(z_{1:n})\pi(z_{1:n})}\right\}$.

S7.3.2 Choice of prior on α

We explain the rationale for the choice of improper prior for α. Suppose at a given iteration of Algorithm 8.1, $\delta = d_n(h_\theta(z_{1:n}), x_{1:n})$ for the current values of θ and $z_{1:n}$. If the prior is $\alpha \sim \text{Gamma}(u, v)$, then the full conditional for α is $q(\alpha|z_{1:n},\theta) = \text{Gamma}(\alpha \mid u,\, v+\delta)$. Thus, the mean and standard deviation of $q(\alpha|z_{1:n},\theta)$ are $u/(v+\delta)$ and $\sqrt{u}/(v+\delta)$, respectively. By setting $v = 0$, we make the distribution of R responsive to the inferred value of δ — specifically, since $\mathbb{E}(R|\alpha) = 1/\alpha$, in expectation R will have the same magnitude as δ/u. In particular, if there is no perturbation and δ is inferred to be very close to 0, then α will be very large and the c-posterior will closely resemble the standard posterior; see Figure 12. To choose an appropriate value of u, we generated calibration data by simulating from the assumed Normal model, and we chose u so that the concentration of the c-posterior is similar to that of the standard posterior when the model is correct.

References

Antoniano-Villalobos, I. and Walker, S. G. Bayesian nonparametric inference for the power likelihood. *Journal of Computational and Graphical Statistics*, 22(4):801–813, 2013.

Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Berger, J. and Berliner, L. M. Robust Bayes and empirical Bayes analysis with ε-contaminated priors. *The Annals of Statistics*, pages 461–486, 1986.

Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York Inc., 1985.

Breiman, L. *Probability*. Addison–Wesley, 1968.

Carota, C., Parmigiani, G., and Polson, N. G. Diagnostic measures for model criticism. *Journal of the American Statistical Association*, 91(434):753–762, 1996.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press, 2006.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2006.

Cox, D. R. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

Dette, H. and Munk, A. Some methodological aspects of validation of models in nonparametric regression. *Statistica Neerlandica*, 57(2):207–244, 2003.

Diaconis, P. and Zabell, S. L. Updating subjective probability. *Journal of the American Statistical Association*, 77(380):822–830, 1982.

Doksum, K. A. and Lo, A. Y. Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, 18(1):443–453, 1990.

Dudley, R. M. *Real Analysis and Probability*. Cambridge University Press, 2002.

Dunson, D. B. and Taylor, J. A. Approximate Bayesian inference for quantiles. *Nonparametric Statistics*, 17(3):385–400, 2005.

Friel, N. and Pettitt, A. N. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 70(3):589–607, 2008.

Geyer, C. J. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, 1991.

Ghosh, S. and Sudderth, E. B. Nonparametric learning for layered segmentation of natural images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2272–2279. IEEE, 2012.

Goutis, C. and Robert, C. P. Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85(1):29–37, 1998.

Gray, R. M. *Entropy and Information Theory*. Springer Science+Business Media, 1990.

Grünwald, P. and van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *arXiv:1412.3730*, 2014.

Hansen, L. P. and Sargent, T. J. Robust control and model uncertainty. *The American Economic Review*, 91(2):60–66, 2001.

Hoff, P. D. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283, 2007.

Holmes, C. C., Bissiri, P. G., and Walker, S. G. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Huber, P. J. *Robust Statistics*. John Wiley & Sons, 2004.

Ibrahim, J. G. and Chen, M.-H. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Jeffrey, R. C. *The Logic of Decision*. McGraw–Hill Book Co. Inc., New York, 1965.

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

Joyce, J. Bayes' theorem. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition, 2008.

Lewis, J. R., MacEachern, S. N., and Lee, Y. Bayesian restricted likelihood methods. *Technical report 878*, 2014.

Li, C., Jiang, W., and Tanner, M. A. General inequalities for Gibbs posterior with nonadditive empirical risk. *Econometric Theory*, 30(06):1247–1271, 2014.

Lindsay, B. and Liu, J. Model assessment tools for a model false world. *Statistical Science*, 24(3):303–318, 2009.

Liu, J. and Lindsay, B. G. Building and using semiparametric tolerance regions for parametric multinomial models. *The Annals of Statistics*, 37(6A):3644–3659, 2009.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

O'Hagan, A. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–138, 1995.

Pettitt, A. Likelihood based inference using signed ranks for matched pairs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45(2):287–296, 1983.

Royall, R. and Tsou, T.-S. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):391–404, 2003.

Rudas, T., Clogg, C. C., and Lindsay, B. G. A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):623–639, 1994.

Smith, J. Q. The multiparameter steady model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 43(2):256–260, 1981.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

Walker, S. and Hjort, N. L. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.

Watson, J. and Holmes, C. Approximate models and robust decisions. *Statistical Science*, 31(4):465–489, 2016.

Whittle, P. *Risk-sensitive Optimal Control*. John Wiley & Sons, Ltd., 1990.

Wilkinson, R. D. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141, 2013.

Zhang, T. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006a.

Zhang, T. Information-theoretic upper and lower bounds for statistical estimation. *Information Theory, IEEE Transactions on*, 52(4):1307–1321, 2006b.