Robust Bayesian inference via coarsening

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

January 29, 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 data exactly, 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 it to a fractional power—thus, inference is often easily implemented with standard methods, 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, nonparametric, power likelihood, relative entropy, tempering.

1 Introduction

In most applications, statistical models are at best idealizations that are known to provide only an approximation to the distribution of the observed data. One might hope that lack of model fit, if sufficiently small, would not significantly impact inferences. Often this does seem to be the case, but sometimes inferences are sensitive to violations of the model assumptions, 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 true data generating process. Although the concepts are general, we focus in particular 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 is focused on robustness to the choice of prior.

Ideally, one would model all aspects of the data generating process correctly, but this is often impractical for a number of reasons. Complex models are more time-consuming to use, more difficult to study theoretically, and less likely to be used in practice due to a general preference for transparent statistical methods in science and other fields. Further, there may be insufficient knowledge of the data generating process to even write down an accurate model. These considerations lead to the following questions. Is it possible to draw valid inferences from a model that may be slightly misspecified? Can this be done in a computationally tractable way? In the context of model averaging and nonparametrics, is there a principled way to be tolerant of models that are not exactly right, but are close enough in some sense?

In this article, we propose a novel approach to robust Bayesian inference that may provide affirmative answers to these questions. Instead of using the standard posterior obtained by conditioning on the event that the observed data are generated by sampling from the model—which is incorrect when the model is misspecified—we condition on the event that the empirical distribution of the observed data is close to the empirical distribution

of data sampled from the model, with respect to some statistical distance on probability measures. We refer to this as a coarsened posterior, or c-posterior, for short.

One can control the type of robustness exhibited by a c-posterior via the choice of statistical distance. For instance, robustness to outliers can be obtained by using a distance that is not strongly affected by moving a small amount of probability mass to an outlying region (e.g., Kolmogorov–Smirnov or 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 distance 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 often 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) establishing the asymptotic form of the c-posterior when certain limits are taken, (3) proving that the c-posterior exhibits robustness to model misspecification (that is, robustness to the form of the likelihood), (4) 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 coarsening, and (5) empirically demonstrating how the power posterior can easily be used to perform robust inference in several key models, using real and simulated data.

The paper is organized as follows. Section 2 describes the c-posterior approach, and Section 3 discusses previous work. Section 4 establishes some theoretical properties concerning asymptotics and robustness. In Section 5, we apply the c-posterior approach to mixture models with an unknown number of components, autoregressive models of unknown order,

and variable selection in linear regression.

2 Method

For now, we assume an i.i.d. setting, but the approach generalizes to time series and regression (Supplement S6). 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 here is that θ_I is the "true" state of nature about which one is interested in making inferences; it may represent some actual underlying truth or may be a useful fiction. Now, suppose there are some unobserved *idealized data* $X_1, \ldots, X_n \in \mathcal{X}$ that are i.i.d. from P_{θ_I}, however, the *observed data* $x_1, \ldots, x_n \in \mathcal{X}$ are actually a slightly corrupted 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 statistical distance $d(\cdot, \cdot)$ 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, and note that due to the corruption, we expect that $P_o \neq P_{\theta_I}$. For intuition, consider the diagram in Figure 1.

If there was no corruption, then we should use the standard posterior—that is, we should condition on the event that $X_{1:n} = x_{1:n}$—however, due to the corruption this would be incorrect. If there was an easy-to-model corrupting process by which $x_{1:n}$ is generated from $X_{1:n}$, then the most sensible approach would be to simply incorporate it into the model—however, this is often impractical, as discussed in the Introduction.

An alternative approach 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 $\pi\big(\theta \mid d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < r\big)$. (For notational specificity, we use π instead of p to denote prior and posterior densities; likewise with Π instead of P for prior and posterior probabilities.) 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 $x_{1:n}$.

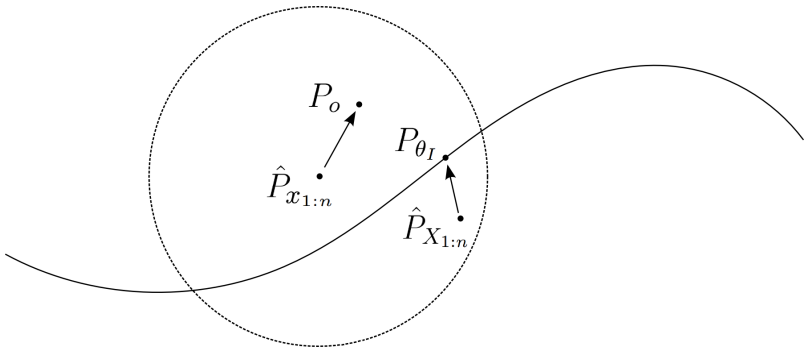

Figure 1: Notional schematic diagram of the idea behind the c-posterior. The ambient space is the set of probability distributions on $\mathcal{X}$, and the curve represents the subset of distributions in the parametrized family $\{P_\theta : \theta \in \Theta\}$. The idealized distribution P_{θ_I} is 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.

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 Section 4.1, we derive the form of the c-posterior as $n \to \infty$. Meanwhile, in Section 4.2, 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 Section 4.3, we show that the c-posterior is robust to changes in P_o that are small with respect to the chosen statistical distance $d(\cdot, \cdot)$. We use the term *statistical distance* broadly to mean any nonnegative function for assessing discrepancy that is meaningful for a given application; it need not be a true metric. 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 distances on 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 applications presented in this paper (see Section 5), we focus 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, etc.). 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 suppose $R \sim \text{Exp}(\alpha)$. 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 4.4, Theorem 4.6), under mild conditions. Empirically we find that the approximation can be quite accurate (see Supplement Figure S1). 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.

A key aspect 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. The choice of $R \sim \text{Exp}(\alpha)$ is not important for robustness (indeed, the theoretical results of Section 4 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.

Section 5 demonstrates several approaches to choosing α; also see Supplement S3.

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 three commonly occurring cases: analytical solution in the case of exponential families with conjugate priors, Gibbs sampling in the case of conditionally conjugate priors, and Metropolis–Hastings MCMC more generally.

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 en-

tropy, whereas usual Bayes factors and model probabilities can be very sensitive to such perturbations for large n (see Supplement S2). In Section 5.2, we apply 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 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—as shown by Equation 2.4—the full conditionals belong to the conjugate family, making them easy to sample from. In Section 5.3, 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. In Section 5.1, we use Metropolis–Hastings for power posterior-based inference in mixtures with a prior on the number of components.

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 Relationships with previous work

The c-posterior is mathematically equivalent to the type of 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 misspecification, while 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; Bissiri et al., 2013). Under certain conditions (see Section 4.1), 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 employed a form of power likelihood obtained by raising the likelihood to a power between 0 and 1. 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), 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). In all such previous research, a fixed power is used, rather than one 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 (2014). 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.

4 Theory

In this section, we establish the asymptotic form of c-posteriors as $n \to \infty$ (Section 4.1), the limit as the distribution of R converges to 0, with n fixed (Section 4.2), and the robustness properties of c-posteriors (Section 4.3). 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.

4.1 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 4.3) and its robustness (Theorem 4.7).

Lemma 4.1. *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 $0 \le \mathbb{1}(\cdot) \le 1$, $|W\mathbb{1}(U_n < V)| \le |W|$, 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 4.2. *Assume* $\mathbb{P}(d(P_{\boldsymbol{\theta}}, P_o) = R) = 0$ *and* $\mathbb{P}(d(P_{\boldsymbol{\theta}}, P_o) < R) > 0$.

Theorem 4.3. *If* $d_n(X_{1:n}, x_{1:n}) \xrightarrow[n\to\infty]{\text{a.s.}} d(P_{\boldsymbol{\theta}}, P_o)$ *and Condition 4.2 is satisfied, then*

$$\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) \xRightarrow[n\to\infty]{} \Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big) \propto G\big(d(P_\theta, P_o)\big)\Pi(d\theta), \tag{4.1}$$

and in fact,

$$\mathbb{E}\big(h(\boldsymbol{\theta}) \mid d_n(X_{1:n}, x_{1:n}) < R\big) \xrightarrow[n\to\infty]{} \mathbb{E}\big(h(\boldsymbol{\theta}) \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big) = \frac{\mathbb{E}h(\boldsymbol{\theta})G\big(d(P_{\boldsymbol{\theta}}, P_o)\big)}{\mathbb{E}G\big(d(P_{\boldsymbol{\theta}}, P_o)\big)} \tag{4.2}$$

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 4.1 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 4.1,

$$\begin{aligned}\mathbb{E}(W \mid U_n < V) \longrightarrow \mathbb{E}(W \mid U < V) &= \frac{\mathbb{E}(W\mathbb{1}(U < V))}{\mathbb{P}(U < V)} \\ &= \frac{\mathbb{E}\big(W\mathbb{E}(\mathbb{1}(U < V)|W, U)\big)}{\mathbb{E}\big(\mathbb{P}(U < V \mid U)\big)} = \frac{\mathbb{E}(WG(U))}{\mathbb{E}G(U)}\end{aligned}$$

since $V \perp U, W$ by construction. This establishes Equation 4.2, and since in particular this holds for any bounded continuous h, Equation 4.1 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 4.3. 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 3. If $R = r_0$ a.s. for some $r_0 > 0$, then $G(r) = \mathbb{1}(r < r_0)$, and by Theorem 4.3 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 4.4. *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 4.2 is satisfied, then Equations 4.1 and 4.2 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 4.5. *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 4.2 is satisfied, then Equations 4.1 and 4.2 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 $\hat{P}_{X_{1:n}} \xRightarrow{\text{a.s.}} P_{\boldsymbol{\theta}}$ 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 4.3 applies. □

4.2 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 4. 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 4.6. *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 $r \to \infty$.*

The proof is in Supplement S7.

4.3 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 Supplement S2).

We show in the following theorem that the asymptotic c-posterior inherits the continuity properties of whatever distance $d(\cdot, \cdot)$ 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 4.7. *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 4.2 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 4.1 with $U = d(P_{\boldsymbol{\theta}}, P_o)$, $U_m = d(P_{\boldsymbol{\theta}}, P_m)$, $V = R$, and $W = h(\boldsymbol{\theta})$. □

Theorem 4.7 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 Supplement S2), 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 distance.

5 Applications

5.1 Mixture models with a prior on the number of components

Consider a finite mixture model, $X_1, \ldots, X_n | k, w, \varphi$ i.i.d. $\sim \sum_{i=1}^{k} w_i f_{\varphi_i}(x)$, and place a prior $\pi(k, w, \varphi)$ on the number of components k, the mixture weights w, and the component parameters φ. This type of model is not robust to misspecification of the family of component distributions $(f_\varphi : \varphi \in \Phi)$. This has negative consequences in practice, since we

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 nice parametric form. We illustrate how the c-posterior enables one to perform inference for the number of components, as well as the mixture weights and the component parameters, in a way that is robust to misspecification of the form of the component distributions. This example also serves to demonstrate the use of Metropolis–Hastings MCMC for inference with a power posterior.

We approximate the relative entropy c-posterior using the power posterior, defined as

$$\pi_c(k, w, \varphi | x_{1:n}) \propto \pi(k, w, \varphi) \prod_{j=1}^{n} \Big(\sum_{i=1}^{k} w_i f_{\varphi_i}(x_j) \Big)^{\zeta_n}.$$

Most approaches to inference in mixture models rely on instantiation of latent variables indicating which component each datapoint comes from, but the power likelihood rules out direct application of such approaches. There are, nonetheless, a few possible approaches to doing inference, for instance, Antoniano-Villalobos and Walker (2013) developed an auxiliary variable technique for mixture power posteriors, or reversible jump MCMC could be used (Green, 1995). To keep things as simple as possible, however, we assume an upper bound on k, say $k \leq m$, and reparametrize the model in a way that enables one to simply use plain-vanilla Metropolis–Hastings (MH) MCMC on a fixed-dimensional space. Specifically, we rewrite the mixture density as $\sum_{i=1}^{m} w_i f_{\varphi_i}(x)$ where $w_i = g(v_i) / \sum_{j=1}^{m} g(v_j)$ for some $v_1, \ldots, v_m > 0$ and $g(v) = \max\{v - c, 0\}$, so that $w_i = 0$ if $v_i \leq c$. Letting $v_1, \ldots, v_m \sim$ Gamma(a, b) i.i.d., conditioning on the event that $\sum_{i=1}^{m} g(v_i) > 0$, and letting $\varphi_1, \ldots, \varphi_m$ be i.i.d. yields a mixture model in which the prior on the number of components (that is, the number of nonzero weights) is $\pi(k) \propto$ Binomial$(k|m, p)\mathbb{1}(k > 0)$ where $p = \mathbb{P}(v_i > c)$. The computation time when doing inference with this setup is approximately 3-4 times that of a standard mixture model algorithm in our experiments, however, in many cases this is compensated for by the improvement in MCMC mixing, so that fewer iterations are needed.

5.1.1 Skew-normal mixture example

To demonstrate robustness to the form of the component distributions, we consider a univariate Gaussian mixture model, applied to data generated i.i.d. from (a) the two-component mixture $\frac{1}{2}\mathcal{SN}(-4,1,5) + \frac{1}{2}\mathcal{SN}(-1,2,5)$, and (b) the four-component mixture $.25\,\mathcal{SN}(-3,.7,-4)+.35\,\mathcal{SN}(0,.8,-3)+.2\,\mathcal{SN}(1,.8,5)+.2\,\mathcal{SN}(3,.5,3)$, where $\mathcal{SN}(\xi,s,a)$ is the skew-normal distribution with location ξ, scale s, and shape a (Azzalini and Capitanio, 1999); see Figure 2 (top). For the model parameters, we assume an upper bound of $m = 15$ components, and define the prior on the component means and precisions as $\mu_i \sim \mathcal{N}(0, 5^2)$ and $\log(\lambda_i) \sim \mathcal{N}(0, 2^2)$ independently for $i = 1, \ldots, m$, where the component densities are of the form $f_{\mu,\lambda}(x) = \mathcal{N}(x|\mu, \lambda^{-1})$. We use a prior on k and w as described above, with $a = 1/m$, $b = 1$, and c such that $p = \mathbb{P}(v_i > c) = 1/m$.

For the c-posterior, we set $\zeta_n = \alpha/(\alpha + n)$ following Equation 2.3, and choose $\alpha = 100$; this can be interpreted as saying that we want the posterior to behave as though at most 100 samples are available, since $n\zeta_n \to \alpha$ as $n \to \infty$. The histograms in the top row of Figure 2 illustrate that based on 100 samples, one can visually determine that there are about two (four, respectively) large groups, and can roughly determine their locations and scales, but cannot determine their precise form—in particular, one cannot tell that they are not actually Gaussian.

For both the two- and four-component examples, for each $n \in \{20, 100, 500, 2000, 10000\}$, five independent datasets of size n were generated, and 10^5 MH sweeps were performed for the standard and coarsened posteriors. Each sweep consists of MH moves on each (μ_i, λ_i) and v_i separately. Rows 2-3 of Figure 2 show the mixture density $\sum_{i=1}^{m} w_i f_{\mu_i,\lambda_i}(x)$ and the individual weighted components $w_i f_{\mu_i,\lambda_i}(x)$ for typical samples from the standard and coarsened posteriors when $n = 10000$. The samples shown are representative, but there is some variability. Rows 4-5 of Figure 2 show the averages of the posteriors on k.

Since the data distribution cannot be represented as a finite mixture of Gaussians, the standard posterior introduces more and more components as n increases, in order to fit the

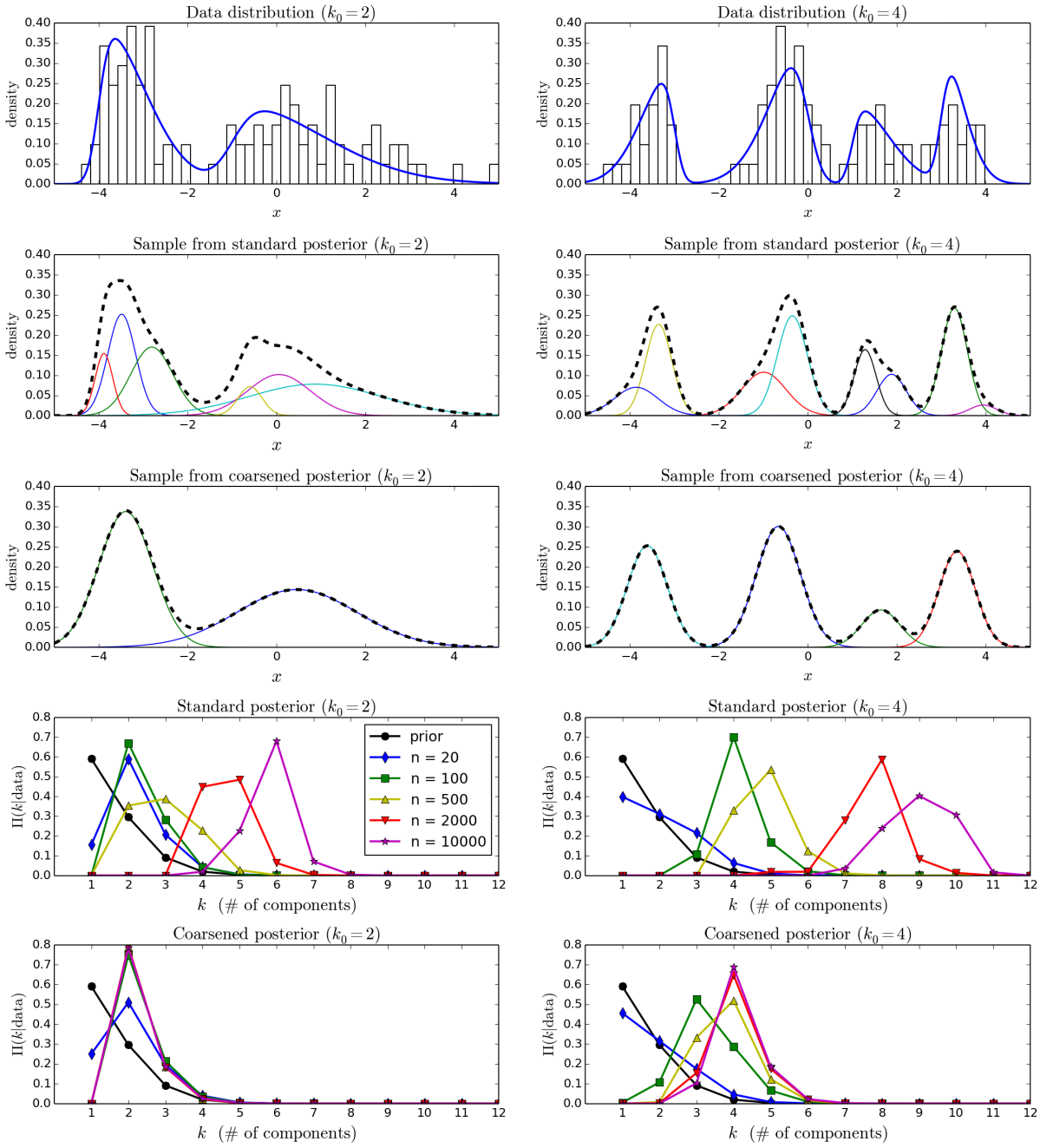

Figure 2: Gaussian mixture with a prior on the number of components k, applied to data from skew-normal mixtures with $k_0 = 2$ and $k_0 = 4$ components. Row 1: Density of the data distribution (blue line) and histogram of $n = 100$ samples. Row 2: Mixture density (dotted black line) and components (solid colors) for typical samples from the posterior when $n = 10^4$. Row 3: Same for the coarsened posterior. Row 4: The standard posterior on k favors larger and larger values as n increases. Row 5: The coarsened posterior on k stabilizes as n increases, favoring the true number of components, $k_0 = 2$ or $k_0 = 4$, in these cases.

data. Meanwhile, in accordance with our visual intuition that, based on 100 samples, there appear to be two (four, resp.) large groups, the $\alpha = 100$ c-posterior shows strong support for two (four, resp.) components, no matter how large n becomes.

Typical samples from the standard posterior provide a better fit to the distribution of the data, however, they have several more components than the true number, obscuring the large-scale structures in the data. Meanwhile, typical samples from the c-posterior do not fit the data distribution as well, but they more closely match the true underlying mixture in terms of the number of components and the weights, locations, and scales of the components. For another example involving mixture models, see Supplement S4.

5.2 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, not only to the form of the distribution of the noise/shocks, but also to misspecification of the structure of the model, such as time-varying noise. This serves as a demonstration of 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 $t \leq 0$ by convention. 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 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,

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

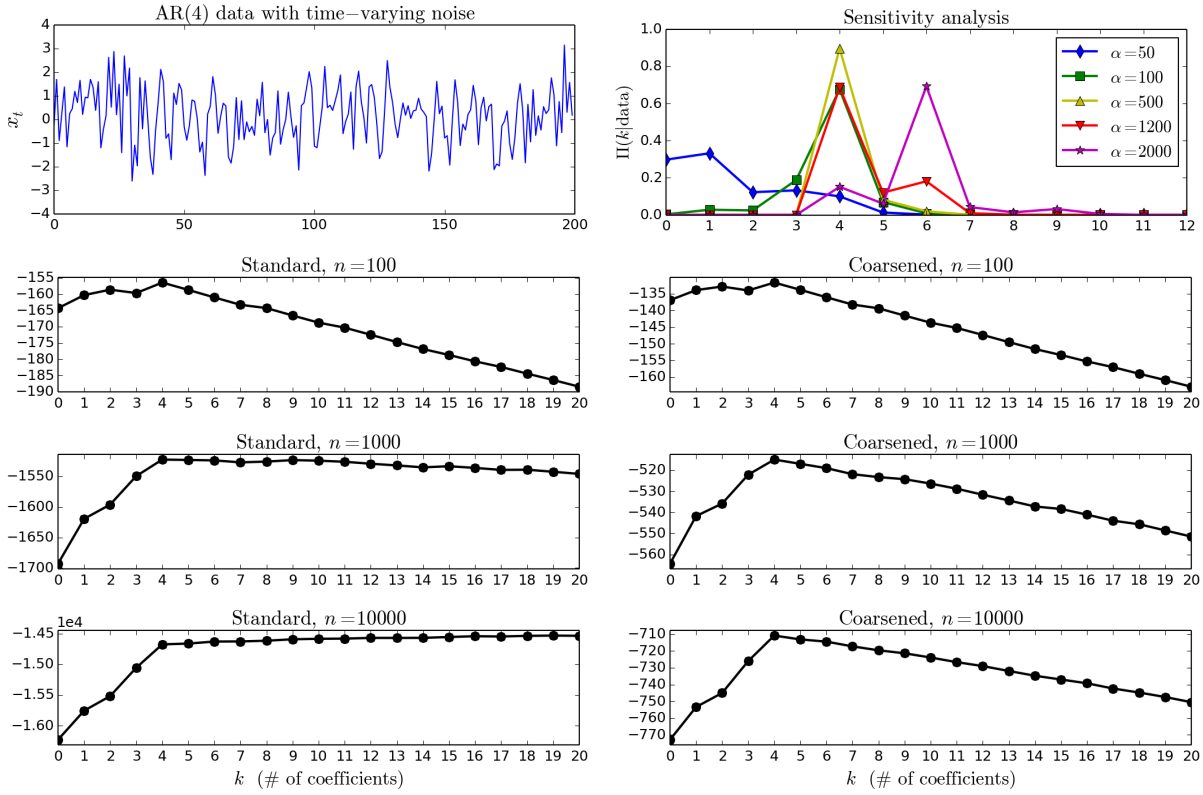

Figure 3: Autoregression example. Upper left: Data sampled from the process in Equation 5.1. Upper right: Sensitivity analysis, displaying the c-posterior on k as α varies, when $n = 10^4$. Lower left: Log marginal likelihood of AR(k) model for $k = 0, 1, \ldots, 20$, on increasing amounts of data from this process. Lower right: Log of coarsened marginal likelihood for the same model, on the same data.

$$= \frac{\exp(\frac{1}{2}\zeta_n^2 v^\intercal \Lambda^{-1} v)}{\sigma_0^k |\Lambda|^{1/2}} \mathcal{N}(x_{1:n} \mid 0, \sigma^2 I_{n\times n})^{\zeta_n}$$

where $\Lambda = \zeta_n M + \sigma_0^{-2} I_{k\times k}$, $M_{ij} = \sum_{t=1}^n x_{t-i}x_{t-j}/\sigma^2$, and $v_i = \sum_{t=1}^n x_t x_{t-i}/\sigma^2$. This, in turn, can be used to compute a robustified posterior on the model order k, defined as $\pi_c(k|x_{1:n}) \propto L_c(k; x_{1:n})\pi(k)$. This is expected to be robust to departures from the AR(k) model that require more than α samples to distinguish, and thus, to favor values of k that are consistent with the data to within this specified tolerance.

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:

$$x_t = \sum_{\ell=1}^{4} \theta_\ell x_{t-\ell} + \varepsilon_t + \tfrac{1}{2}\sin t \tag{5.1}$$

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 $t \leq 0$. 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, and take $\sigma_0^2 = 1$. If one expects a particular amount of misspecification, then a principled choice of α can be made, e.g., see Section 5.3. Here, we assess sensitivity to the choice of α by considering the c-posterior on k as α varies, when $n = 10^4$, with a weakly informative Geometric(0.1) prior on k (i.e., $\pi(k) = 0.9^k 0.1$); see Figure 3 (upper right). We see that there is a fairly wide range of values that give similar results: for α's between 100 and 1200, the large majority of the mass is on the correct value of k, namely $k = 4$.

To visualize what happens as n increases, we set $\alpha = 500$, and consider the log of the marginal likelihood. Due to the misspecification, the standard posterior strongly favors values of k much greater than 4 when n gets sufficiently large; see Figure 3 (lower left) and note the scale. Meanwhile, the c-posterior stabilizes to a distribution on k favoring $k = 4$; see Figure 3 (lower right). For values of n less than α, the standard and coarsened approaches yield similar results, however, as n increases beyond α, they differ markedly.

More generally, this pattern is typical of the log marginal likelihood when comparing models of increasing complexity. The log marginal likelihood penalizes more complex models 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 S2), 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, by behaving as though the sample size is no larger than α.

5.3 Variable selection in linear regression

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

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

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 linearity assumed in the mean function $\beta^{\mathsf{T}} x_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. For instance, if the mean is actually $\beta_1 g(x_{i1})$ where g is close to but not exactly linear, and x_{i1} is correlated with other covariates, then the posterior will typically make additional coefficients nonzero in order to compensate.

We demonstrate how the c-posterior provides robustness to misspecification of this type. This example also provides an opportunity to show 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 $\pi_c(\beta, \lambda|y) \propto p(y|\beta,\lambda)^{\zeta_n}\pi(\beta,\lambda)$. 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. We give the formulas here without justification:

$$\pi_c(\lambda|\beta, y) = \text{Gamma}\left(\lambda \,\middle|\, a + \tfrac{1}{2}n\zeta_n,\, b + \tfrac{1}{2}\zeta_n \textstyle\sum_{i=1}^n (y_i - \beta^{\mathsf{T}} x_i)^2\right)$$

and one can sample from $\pi_c(\beta_j|\beta_{-j}, \lambda, y)$, where $\beta_{-j} = (\beta_\ell : \ell \neq j)$, by setting $\beta_j = 0$ with probability

$$\Pi_c(\beta_j = 0 \mid \beta_{-j}, \lambda, y) = \left(1 + \sqrt{L_0/L}\,\exp\left(\tfrac{1}{2}LM^2\right)\frac{r + \sum_{\ell\neq j}\mathbb{1}(\beta_\ell \neq 0)}{s + \sum_{\ell\neq j}\mathbb{1}(\beta_\ell = 0)}\right)^{-1}$$

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 $\delta_i = y_i - \sum_{\ell\neq j}\beta_\ell x_{i\ell}$, and otherwise sampling β_j from $\mathcal{N}(M, L^{-1})$.

5.3.1 Simulation example

Consider a simulation example where the mean of the observed data is a slightly nonlinear function of a single covariate, plus a constant offset:

$$y_i = \beta_{01} + \beta_{02}(x_{i2} + \tfrac{1}{16}x_{i2}^2) + \varepsilon_i \tag{5.2}$$

where $\beta_{01} = -1$, $\beta_{02} = 4$, and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0,1)$. Following standard practice, suppose $x_{i1} = 1$, to accommodate a constant offset. 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 $\tilde{X}_i \sim \mathcal{SN}_5(\Omega, a)$ with shape $a = (0.6, 2.7, -3.3, -4.9, -2.5)$ 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 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 4 (top left) shows a scatterplot of y_i versus x_{i2} for 200 samples, as well as the mean as a function of x_{i2}.

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 $a = b = 1$. To choose α, note that the c-posterior is obtained by conditioning on the conditional relative entropy estimate being less than R, where $R \sim \text{Exp}(\alpha)$ (see 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 misspecification/contamination to shift the mean function by approximately $\pm\delta$ on average, and the noise has standard deviation σ, then it is reasonable to choose α so that $\mathbb{E}R \approx \delta^2/(2\sigma^2)$, i.e., $\alpha \approx 2\sigma^2/\delta^2$. In the present situation, by cheating and using our knowledge of the truth, we choose $\delta = 0.2$ and $\sigma = 1$, leading to $\alpha = 50$.

For each $n \in \{100, 1000, 5000, 10000, 50000\}$, ten datasets were generated, and for both the standard posterior and the coarsened posterior, 50000 Gibbs sweeps were performed on each dataset, the first 5000 of which were discarded as burn-in.

Figure 4 (middle) shows the average of these posteriors on k over the 10 datasets, for the standard and coarsened posteriors. Note that the "true" number of nonzero coefficients in Equation 5.2 is $k = 2$ (β_{01} and β_{02}).

Figure 4 (bottom) shows the posterior cumulative distribution function (c.d.f.) and 95% credible interval for each coefficient $\beta_1, \ldots, \beta_6$ when $n = 10000$, for the standard and coarsened posteriors. Recall that the "true" values are $\beta_1 = -1$, $\beta_2 = 4$, and $\beta_3 = \cdots = \beta_6 = 0$. The 95% intervals for the standard posterior are quite far from the true values of β_1, β_4, and β_5, while 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 in addition to incorrectly inferring which coefficients are nonzero, the standard posterior can also lead to incorrect inferences about the values of the nonzero coefficients. The c-posterior mitigates this by more appropriately calibrating the

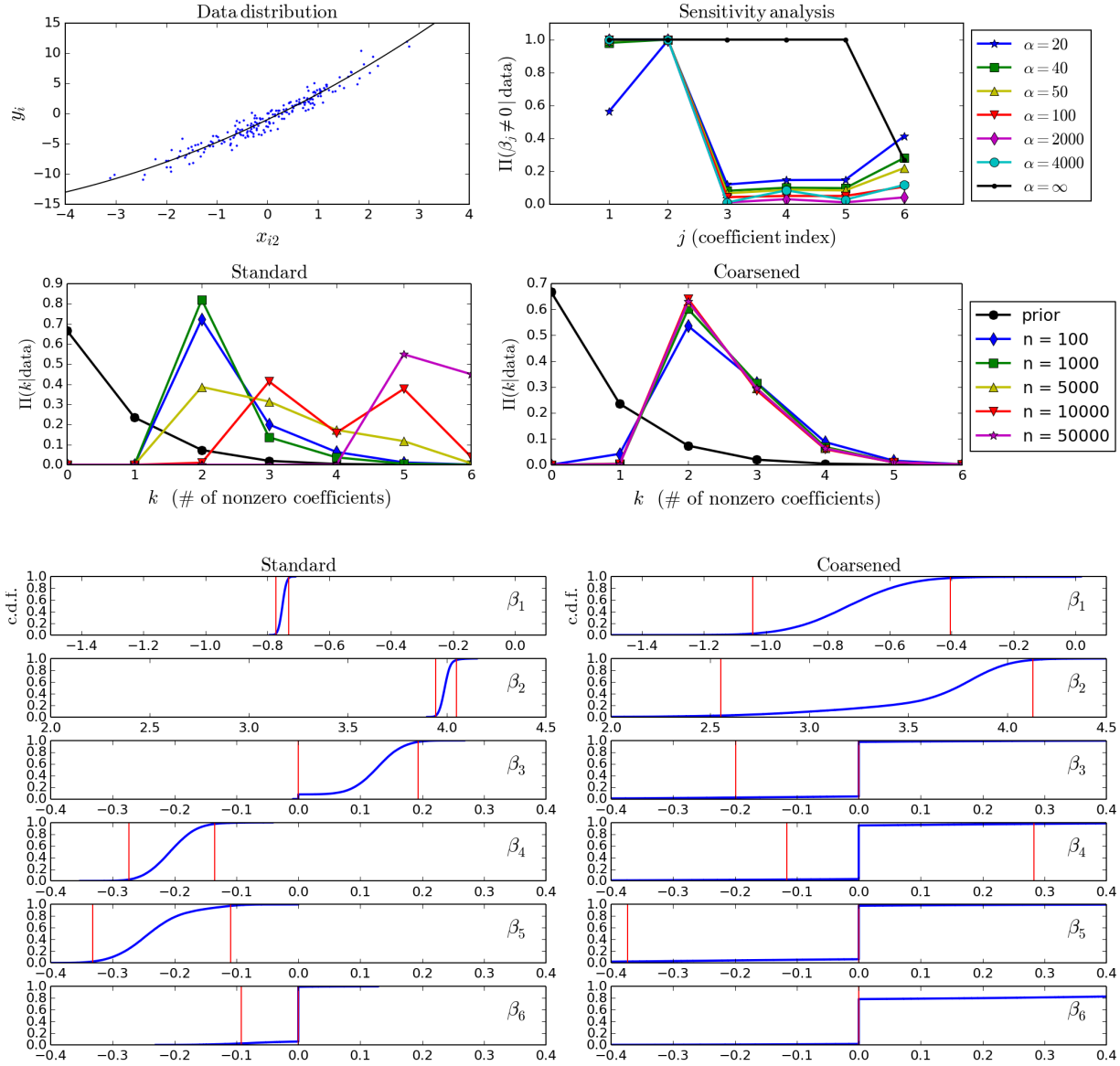

Figure 4: Variable selection with simulated data. Top left: Scatterplot of the target variable y_i versus x_{i2}, as well as the mean function (black line). Top right: Sensitivity analysis displaying the posterior inclusion probabilities as α varies, when $n = 50000$. Middle left: The standard posterior on the number of nonzero coefficients k favors larger values as n increases. Middle right: The c-posterior favors the "true" number, $k = 2$, even as n grows. Bottom left: Posterior c.d.f. for each coefficient (blue), and 95% credible interval (red). Bottom right: Same, for the c-posterior.

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.

To address the question of sensitivity to the choice of α, Figure 4 (top right) shows the posterior inclusion probabilities for each coefficient as α varies; the probabilities shown are averaged over three datasets with $n = 50000$. Over a very wide range of α values (from around 40 to as high as 4000 or so), we obtain similar results — namely, the two correct coefficients (β_1 and β_2) are identified as being nonzero with very high probability, and the rest have low inclusion probability. On the other hand, as α varies, the posterior credible intervals shrink or expand, affecting coverage.

5.3.2 Modeling birthweight of infants

The Collaborative Perinatal Project (CPP) collected data from a large sample of mothers and their children, measuring many medical and socioeconomic variables from before and during pregnancy, as well as early childhood (Klebanoff, 2009). Using data from a follow-up study that collected additional information for a subset of the CPP participants, we illustrate how the c-posterior can be used to analyze the relationship between birthweight and a number of predictor variables.

The dataset we use contains $n = 2379$ subjects and 71 covariates that are potentially predictive of birthweight. The data are preprocessed to normalize each covariate as well as the target variable, by subtracting off the sample mean and dividing by the sample standard deviation for each. As usual, a constant covariate is appended, making $p = 72$. We use the same prior parameters as in the simulation example. Rather than choose a single value of α, we explore the data at varying levels of coarseness, by considering a range of α values.

For each $\alpha \in \{100, 500, 1000, 2000, \infty\}$, we run the sampler for 10^4 Gibbs sweeps, discarding the first 1000 sweeps as burn-in. The sampler is initialized by setting all the coefficients to zero; initializing with a sample from the prior yields identical results. To interpret α in terms of Euclidean notions, we estimate from posterior samples that $\lambda \approx 2.5$

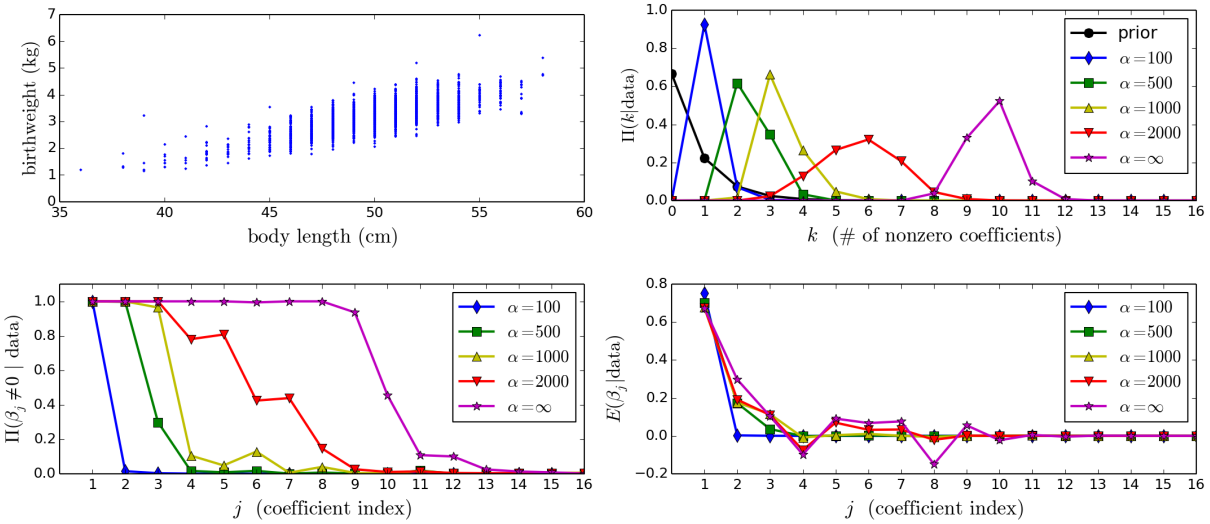

Figure 5: Variable selection for modeling birthweight. Upper left: Scatterplot of birthweight (the target variable) versus body length at birth. Upper right: The standard posterior includes several more nonzero coefficients than the c-posteriors. Lower left: Posterior probability of inclusion for each coefficient; only the top 16 are shown, see list below. Lower right: Posterior mean of each coefficient, for the same 16. (Top 16 variables: 1. Body length, 2. Mother's weight at delivery, 3. Gestation time, 4. African-American, 5. Center 6, 6. Center 2, 7. Center 3, 8. Mother's weight prepregnancy, 9. Previously pregnant, 10. Cigarettes per day, 11. # prenatal checkups, 12. Smoker/non-smoker, 13. Mother's BMI prepregnancy, 14. # previous pregnancies, 15. Triglyceride level, 16. Center 10.)

to 3, and thus, by the formula $\alpha \approx 2\sigma^2/\delta^2 = 2/(\lambda\delta^2)$ derived in Section 5.3.1, the values of α above roughly correspond to allowing for misspecification/contamination of magnitude $\delta \in \{0.09, 0.04, 0.03, 0.02, 0\}$, respectively, or, when scaled to the original units, roughly $\delta_{\text{kg}} \in \{0.045, 0.02, 0.015, 0.01, 0\}$ kilograms.

The posterior on the number of nonzero coefficients k (Figure 5, upper right) and the posterior probability of inclusion (Figure 5, lower left) show that the standard posterior includes around 10 out of the 72 coefficients, while the c-posterior employs a more parsimonious representation, depending on α. At $\alpha = 100$ ($\delta_{\text{kg}} \approx 0.045$), typically only a single variable is included, namely, body length. It makes sense that body length would be

strongly predictive of weight, and the scatterplot in Figure 5 (upper left) confirms this. At $\alpha = 500$ ($\delta_{\text{kg}} \approx 0.02$), both body length and mother's weight at delivery are included, as well as gestation time with somewhat lower probability; again, it makes sense for these to be predictive of birthweight. As α increases, additional variables are included to account for finer aspects of the data, until we reach the standard posterior at $\alpha = \infty$.

All of the variables included by the standard posterior could conceivably be predictive of birthweight, although after adjusting for primary variables such as body length, it is possible that they are only being included due to misspecification. Since it seems likely that there would be misspecification at least at the $\delta_{\text{kg}} = 0.01$ kg level (i.e., ≈ 0.02 pounds, $\alpha = 2000$), the high probability placed on some of the additional variables included by the standard posterior is dubious, as is the precision with which it purports to infer the corresponding coefficients. Further, it seems inevitable that if n were larger, then even more coefficients would be included by the standard posterior. If there is misspecification, then as n grows, eventually the interpretation of which coefficients are included becomes less related to practically significant associations and more related to the fact that the model is compensating for its limitations.

6 Conclusion

The c-posterior approach seems promising as a general method of robust Bayesian inference. There are a number of directions that would be interesting to pursue in future work. It would be useful to have a tractable way of inferring α from data, since this would enable one to assess the amount of misspecification, and ideally, to achieve statistical efficiency when the model is correct. Further investigation of the accuracy of the power posterior approximation is needed, both theoretically and empirically. We have focused on relative entropy due to the computational advantages, but it would be interesting to explore using other statistical distances, particularly if fast inference methods could be developed for them as well. It would be beneficial if precise guarantees could be provided regarding

frequentist coverage properties of the c-posterior, under misspecification. Finally, the scope of application of this approach is not limited to Bayesian inference; it would be interesting to explore adaptations to frequentist procedures.

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.

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.

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.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. A general framework for updating belief distributions. *arXiv:1306.6430*, 2013.

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.

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.

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. *Interface Foundation of North America*, 1991.

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.

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.

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.

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.

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.

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

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.

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. *arXiv preprint arXiv:1402.6118*, 2014.

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.

Supplementary material for "Robust Bayesian inference via coarsening"

S1 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. The c-posterior inherits the continuity properties of the chosen statistical distance, and thus, exhibits robustness to small departures from the model—that is, small changes to the data distribution result in small changes to the c-posterior. Asymptotically, the c-posterior takes a relatively simple form, facilitating computation and analysis.

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. In particular, this occurs if the amount of misspecification is less than expected. This problem is illustrated by the overly wide c-posterior credible interval for β_2 in the variable selection example shown in Figure 4. On the other hand, the same example shows that in the presence of misspecification, c-posterior credible intervals tend to behave better than standard credible intervals in terms of containing the true value. 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 misspecification 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 model misspecification, 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 (as discussed in the Introduction), 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 misspecification problem, 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, and it will 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.

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 for a broad range of misspecification issues, the distribution of $x_{1:n}|X_{1:n}$ needs to allow for distributional 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.

Strategies for choosing α

Choosing α requires some insight and effort. However, choosing a model already requires a significant amount of insight and domain knowledge. If one has enough *a priori* knowledge to design an appropriate model, then it seems reasonable to expect that one could make an appropriate choice of α, based on his or her confidence in the accuracy of that model.

The examples in this article illustrate three strategies for choosing α:

- Strategy #1: Set the mean neighborhood size $\mathbb{E}R = 1/\alpha$ to match the amount of misspecification expected. To help quantify *a priori* knowledge in terms of neighbor-

hood size R, it is possible in some cases to roughly translate intuitive notions like Euclidean distance into relative entropy. (See the variable selection examples and the toy Bernoulli example.)

- Strategy #2: Rule of thumb – to be robust to perturbations that would require at least N samples to distinguish, set $\alpha = N$. 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. This strategy can be used to translate intuitions about the amount of information in a sample of size N into an appropriate choice of α. (See the skew-normal mixture example.)

- Strategy #3: Consider a range of α values, for sensitivity analysis or exploratory analysis. (See the autoregression example, the birthweight regression example, and the Shapley galaxy mixture example.)

Inferring α based on the data

It would be nice to have a way of choosing α based on the data. In principle, one could compute a nonparametric density estimate of the observed data distribution, compute the distance to the estimated model, and choose α based on this distance — but this would be computationally intensive and thus not very practical.

It is natural to think that one could treat α like a hyperparameter, however, this has some fundamental issues. It is not clear whether one could choose α using standard techniques such as maximizing the marginal likelihood, because α is not a model parameter in the usual sense. The reason is that the c-posterior does not arise from specifying a model on the observed data — in other words, there is no such thing as $p(x_{1:n}|\alpha)$ when using coarsening. One might think that a natural choice of "marginal likelihood" for α would be

$$\mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \alpha\big) = \int_\Theta \mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \theta, \alpha\big)\Pi(d\theta),$$

but this is always maximized by sending α to 0 (i.e., making the distribution of R diverge to ∞).

Alternatively, in the convolution representation in Equation 2.1, one might be able to view the "kernel" $G(d_n(x'_{1:n}, x_{1:n})) = \exp(-\alpha d_n(x'_{1:n}, x_{1:n}))$ as an unnormalized density for $x_{1:n}$ given $x'_{1:n}$, and define $p(x_{1:n}|\alpha)$ to be

$$\int_\Theta \int_{\mathcal{X}^n} \frac{\exp(-\alpha d_n(x'_{1:n}, x_{1:n}))}{Z(x'_{1:n}, \alpha)} P_\theta^n(dx'_{1:n})\Pi(d\theta).$$

However, it is not at all clear whether this will be computationally tractable, even approximately.

How to efficiently choose α based on the data remains an open question, and is a topic for future research.

What type of deviations does coarsening tolerate?

The type of deviations tolerated by the c-posterior depends on the choice of statistical distance $d_n(\cdot, \cdot)$ between distributions, and the size of deviation tolerated is governed by the distribution of 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 where the misspecification is small with respect to relative entropy, in order to illustrate robustness of the relative entropy c-posterior to such perturbations. One could choose any misspecification that is small with respect to relative entropy, and the relative entropy c-posterior (as well as the power posterior) would be robust to it.

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 a straightforward matter 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 $\prod_{j=1}^{n} \left(\sum_{i=1}^{k} w_i f_{\varphi_i}(x_j) \right)^{\zeta_n}$, and it seems that introducing $z_1, \ldots, z_n$ no longer leads to nice full conditionals. On the other hand, it may be possible to use a different set of latent variables; see Antoniano-Villalobos and Walker (2013) for the case of mixtures.

S2 Lack of robustness of the standard posterior

Standard model selection methods do not address the model misspecification 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 S2.1), while c-posteriors are robust to small changes in P_o (as shown in Section 4.3). 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 $n \to \infty$ (see Section 4).

S2.1 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 $\theta_k^n = \operatorname{argmax}_{\theta\in\Theta_k} p(x_{1:n}|\theta)$ is the maximum likelihood estimator for model k, $\theta_k^* = \operatorname{argmin}_{\theta\in\Theta_k} D(p_o\|p_\theta)$ is the minimal Kullback–Leibler (KL) point within model k, and $H(\theta; p_o) = -\int p_o\big(\nabla_\theta^2 \log p_\theta\big)$. Here, $a_n \sim b_n$ means $a_n/b_n \to 1$. Letting $f_n(k) = -\frac{1}{n}\log p(x_{1:n}|\theta_k^n)$, this implies that

$$p(x_{1:n}|k) \sim c_k e^{-nf_n(k)} n^{-t_k/2} \tag{S2.1}$$

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.

S3 Toy example: Bernoulli trials

This toy example is intended to serve as a self-contained explanation of the method in the simplest possible setting, and also 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. Of course, in such a simple setting as Bernoulli trials, it would be easy to improve the model to account for issues such as these. However, for more complex models it is often not so easy, and we need a method that works well even with complex models. To reiterate, we are not suggesting that the c-posterior is the best approach to the Bernoulli problem—the purpose of this toy example is simply to provide a concrete illustration of how the method works, in a setting that is easy to understand.

Suppose it is known that the corruption affects the distribution of the data by only a small amount. We can formulate this mathematically by considering $X_{1:n}$ to represent some hypothetical idealized data which do follow the model, and supposing that the observations $x_{1:n}$ are close to the idealized data in some distributional sense, but not necessarily equal to them. A natural way to define distributional "closeness" is in terms of the relative entropy

$D(\hat{p}_x \| \hat{p}_X) = \sum_{i=0}^{1} \hat{p}_x(i) \log(\hat{p}_x(i)/\hat{p}_X(i))$ between the empirical distributions of $x_{1:n}$ and $X_{1:n}$, i.e., $\hat{p}_x(1) = \bar{x}$ and $\hat{p}_x(0) = 1 - \bar{x}$ in this example.

Due to the corruption, it is inappropriate to condition on the idealized data $X_{1:n}$ being exactly equal to the observed data $x_{1:n}$. Instead, if it is known that $X_{1:n}$ is close to $x_{1:n}$ in the sense that $D(\hat{p}_x \| \hat{p}_X) < r$, and nothing more is known about the nature of the corruption, then a natural Bayesian approach would be to condition on the event that $D(\hat{p}_x \| \hat{p}_X) < r$, that is, to use $\Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x || \hat{p}_X) < r\big)$ instead of $\Pi(\mathrm{H}_0 | x_{1:n})$.

In practice, one will typically only have a rough idea about the amount of corruption, and thus, it makes sense to put a prior on r, say, $R \sim \mathrm{Exp}(\alpha)$. This leads us to consider the following coarsened posterior (c-posterior) for inferences about H_0 and H_1:

$$\Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x || \hat{p}_X) < R\big). \tag{S3.1}$$

In other words, we consider $\Pi(\mathrm{H}_0 | Z = 1)$ where $Z = \mathbb{1}(D(\hat{p}_x || \hat{p}_X) < R)$. How should we choose α? In this example, we can interpret the neighborhood size r in terms of intuitive Euclidean notions by using 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 $D(\hat{p}_x || \hat{p}_X) \approx 2|\bar{x} - \bar{X}|^2$, and thus, if we expect the corruption to 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 S3.1 can be computed exactly (see Section S3.1 for details), however, in more complex cases, an approximation is needed. The power likelihood approximation from Section 2.1, when applied to this example, yields

$$\Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x || \hat{p}_X) < R\big) \approx 1/\big(1 + 2^{\alpha_n} B(1 + \alpha_n \bar{x},\, 1 + \alpha_n(1 - \bar{x}))\big) \tag{S3.2}$$

where $\alpha_n = 1/(1/n + 1/\alpha)$ and $B(a, b)$ is the beta function (see Section S3.1 for details). Comparing this to the standard posterior,

$$\Pi\big(\mathrm{H}_0 \,\big|\, X_{1:n} = x_{1:n}\big) = 1/\big(1 + 2^n B(1 + n\bar{x},\, 1 + n(1 - \bar{x}))\big), \tag{S3.3}$$

note that the only difference is that n has been replaced by α_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, and the data are corrupted in such a way that $x_1, \ldots, x_n$ behave like i.i.d. samples from Bernoulli(0.51). Figure S1 (top and middle) shows the probability of H_0 under the standard posterior, the exact c-posterior, and the approximate c-posterior (Equations S3.3, S3.1, and S3.2, 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.) Meanwhile, 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. Note, further, that the curve for the approximate c-posterior is directly on top of the curve for the exact c-posterior—the approximation is so close that the two are 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 S1 (bottom) shows the three posteriors on data $x_1, \ldots, x_n$ i.i.d. $\sim$ Bernoulli(0.56). We see that in this case, the c-posterior behaves more like the standard posterior, favoring H_1 when n is sufficiently large.

It is important to note that, unlike the standard posterior, the c-posterior does not concentrate as $n \to \infty$. This is appropriate, since in the presence of corruption, some uncertainty always remains about the true distribution, no matter how many data points are observed.

To reiterate, we are not suggesting that the c-posterior is the best approach to the corrupted Bernoulli problem. In such a trivial setting as this, there are obvious alternatives that would be preferable, such as an interval null, $\mathrm{H}_0 : 1/2 - \varepsilon < \theta < 1/2 + \varepsilon$. The purpose

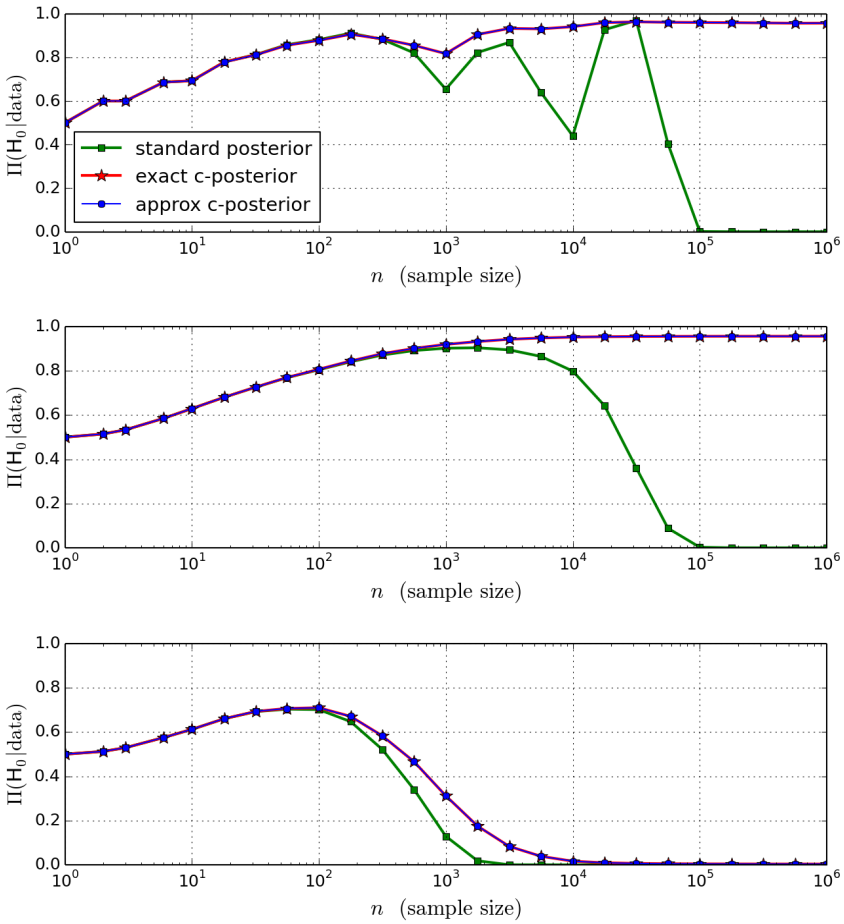

Figure S1: Bernoulli trials example. Top: Results from a single sequence $x_1, x_2, \ldots$ i.i.d. $\sim$ Bernoulli(0.51). Middle: Average over 1000 sequences $x_1, x_2, \ldots$ i.i.d. $\sim$ Bernoulli(0.51). Bottom: Same as middle, but with 0.56 instead of 0.51. In all three plots, the approximate c-posterior is indistinguishable from the exact c-posterior.

of this toy example is simply to provide a concrete illustration of how the method works, in a setting that is easy to understand.

S3.1 Derivations for toy Bernoulli example

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

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

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 $S|\mathrm{H}_0 \sim \mathrm{Binomial}(n, 1/2)$ and $S|\mathrm{H}_1 \sim \mathrm{BetaBinomial}(n, 1, 1) = \mathrm{Uniform}\{0, 1, \ldots, n\}$. To derive the approximation in Equation S3.1, we use Equation S5.2:

$$\mathbb{E}\big(\exp(-\alpha D(\hat{p}_x||\hat{p}_X)) \,\big|\, \theta, h\big) \approx \sqrt{\alpha_n/\alpha}\exp(-\alpha_n D(\hat{p}_x||p_\theta)) = c\prod_{i=1}^{n} p_\theta(x_i)^{\alpha_n/n}$$

where $\alpha_n = 1/(1/n + 1/\alpha)$, $p_\theta(x) = \mathrm{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x \in \{0, 1\}$, and c is a constant that does not depend on θ or h. If $h = \mathrm{H}_1$, then this yields

$$\begin{aligned}\mathbb{P}(Z = 1|\mathrm{H}_1) &= \mathbb{E}\Big(\mathbb{E}\big(\exp(-\alpha D(\hat{p}_x||\hat{p}_X)) \,\big|\, \theta, \mathrm{H}_1\big) \,\Big|\, \mathrm{H}_1\Big) \approx \mathbb{E}\Big(c\prod_{i=1}^{n} p_\theta(x_i)^{\alpha_n/n} \,\Big|\, \mathrm{H}_1\Big) \\ &= c\int_0^1 \theta^{\alpha_n\bar{x}}(1-\theta)^{\alpha_n(1-\bar{x})}\,d\theta = cB\big(1 + \alpha_n\bar{x},\, 1 + \alpha_n(1-\bar{x})\big).\end{aligned}$$

If $h = \mathrm{H}_0$, then $\theta = 1/2$ with probability 1, so $\mathbb{P}(Z = 1|\mathrm{H}_0) \approx c(1/2^{\alpha_n/n})^n = c/2^{\alpha_n}$. Thus, $\Pi(\mathrm{H}_0|Z = 1) = \mathbb{P}(Z = 1|\mathrm{H}_0)/(\mathbb{P}(Z = 1|\mathrm{H}_0) + \mathbb{P}(Z = 1|\mathrm{H}_1)) \approx 1/(1 + 2^{\alpha_n}B(1 + \alpha_n\bar{x},\, 1 + \alpha_n(1-\bar{x}))$.

S4 Mixture example with Shapley galaxy dataset

The galaxy dataset of Roeder (1990) is a classic benchmark for nonparametric mixture models, but it is somewhat outdated, and rather small, with only $n = 82$ galaxies. Drinkwater et al. (2004) provide a more recent and larger dataset of the same type, consisting of the velocities of 4215 galaxies in the Shapley supercluster, a large concentration of gravitationally interacting galaxies; see Figure S2. The clustering tendency of galaxies continues to be a subject of interest in astronomy. However, due to the filament-like nature of the distribution of galaxies, it seems likely that any such clusters will not be Gaussian.

Nonetheless, with the c-posterior approach, a Gaussian mixture model can be used to good effect to identify clusters that are approximately normal. By varying α, one can explore the data at varying levels of precision, allowing for greater or smaller departures from normality.

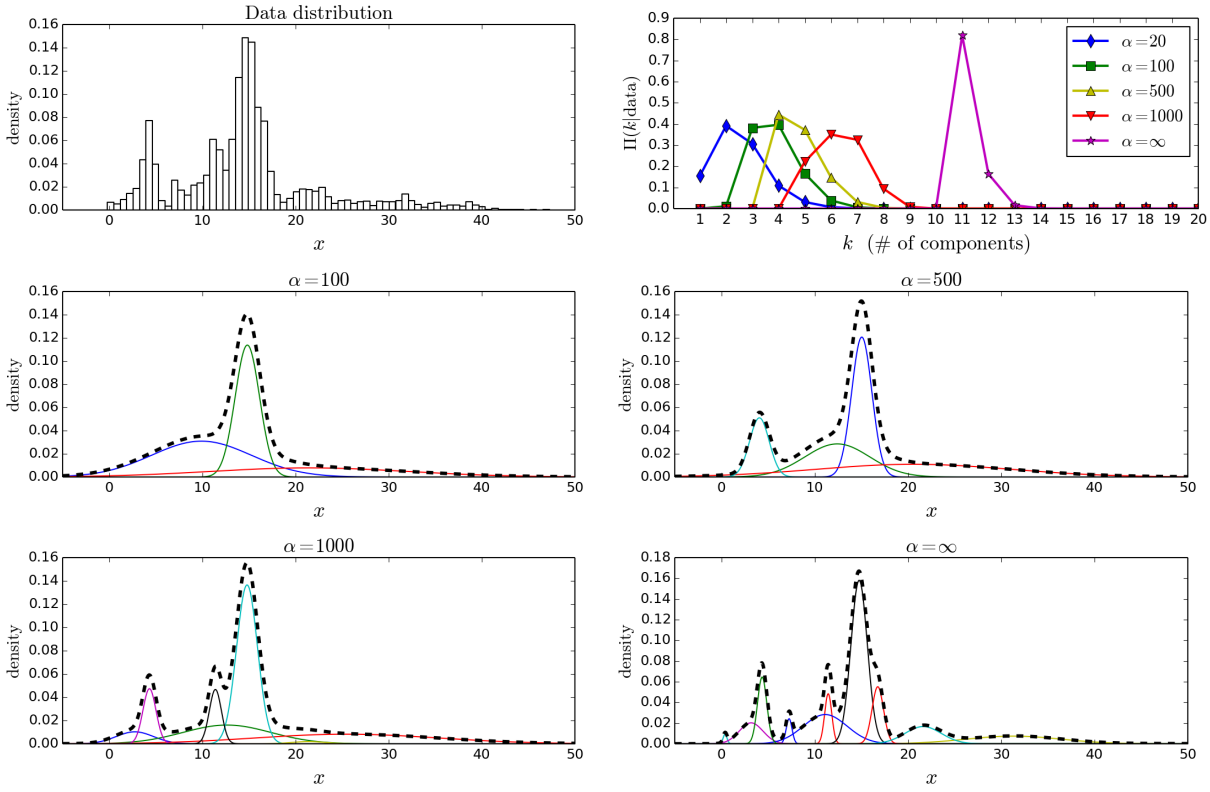

Figure S2: Gaussian mixture with a prior on the number of components k, applied to the Shapley galaxy data. Top left: Histogram of the data, in units of 1,000 km/s, excluding a small amount of data extending in a tail up to 80,000 km/s. Top right: C-posterior on k for a range of α values; $\alpha = \infty$ is the standard posterior. Middle and bottom: Mixture density (dotted black line) and components (solid colors) for prototypical samples from the c-posterior, for a range of α values.

We use the same model as in Section 5.1.1, but with $m = 20$ and a data-dependent choice of prior parameters: $\mu_i \sim \mathcal{N}(\bar{x}, \hat{\sigma}^2)$ and $\log(\lambda_i) \sim \mathcal{N}(\log(4/\hat{\sigma}^2), 2^2)$. For $\alpha \in \{20, 100, 500, 1000\}$, we run the sampler for 10^5 MH sweeps, with a burn-in of 10^4 sweeps. For the standard posterior ($\alpha = \infty$), mixing is considerably slower; to compensate, we use 10^6 sweeps with a burn-in of 2.5×10^5. This illustrates how inference can be easier under the c-posterior.

As shown in Figure S2, when α is small, the c-posterior tolerates greater departures from normality, and uses a smaller number of components to represent the data. For instance, from a glance at the histogram, one can visually distinguish three or four large

groups which appear roughly unimodal, and when α is around 100 to 500, samples from the c-posterior tend to provide a mixture representation that corresponds well to these intuitive groups. For larger values of α, additional mixture components are employed, to account for finer and finer grained aspects of the data distribution. By the time $\alpha = \infty$, i.e., the standard posterior, the large-scale structures have mostly been fragmented into many small components.

Of course, in a univariate setting like this, one can already visually see the large-scale groups, but clusters in high-dimensional data are not so easy to visualize, and having a tool like the c-posterior to find structures at varying levels of precision may be very useful. In most applications, the primary use of mixture models is not density estimation, but rather, to provide an interpretable summary of the data in terms of clusters, and in these cases the c-posterior approach may have much to offer.

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 4.1, 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

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

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 4.6, 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,

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

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,

$$D(p \| q) \approx \tfrac{1}{2}\chi^2(p, q) = \tfrac{1}{2}(p' - q')^\intercal C(q')^{-1}(p' - q') \qquad \text{(S5.3)}$$

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 $R \sim$

Exp(α). Then by Lemma 4.1, 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

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

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 5.2, 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 contamination/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 corrupted 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 4.1, 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 5.3, we apply this to perform robust variable selection in linear regression.

S7 Proof of small-sample behavior

Proof of Theorem 4.6. (Discrete case) By the dominated convergence theorem,

$$\frac{1/G(0)}{|\mathcal{E}(x_{1:n})|} \, \mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R/\alpha \,\big|\, \theta\big) = \frac{1/G(0)}{|\mathcal{E}(x_{1:n})|} \sum_{x'_{1:n} \in \mathcal{X}^n} G(\alpha d_n(x'_{1:n}, x_{1:n})) \prod_{i=1}^n p_\theta(x'_i)$$

$$\xrightarrow[\alpha\to\infty]{} \frac{1/G(0)}{|\mathcal{E}(x_{1:n})|} \sum_{x'_{1:n}\in\mathcal{E}(x_{1:n})} G(0)\prod_{i=1}^n p_\theta(x'_i) = \prod_{i=1}^n p_\theta(x_i).$$

(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

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

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 S7.1 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 S7.1. *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 $x \in B_\delta$, $h(x_0) - \varepsilon \leq h(x) \leq h(x_0) + \varepsilon$. 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,

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

□

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.

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

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.

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

Drinkwater, M. J., Parker, Q. A., Proust, D., Slezak, E., and Quintana, H. The large scale distribution of galaxies in the Shapley supercluster. *Publications of the Astronomical Society of Australia*, 21(1):89–96, 2004.

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

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

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

Roeder, K. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.