ROBUST BAYESIAN INFERENCE VIA COARSENING

JEFFREY W. MILLER AND DAVID B. DUNSON

Duke University, Department of Statistical Science

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 simple, coherent approach to Bayesian inference that improves robustness to perturbations from the model: rather than condition on the data exactly, one conditions on a neighborhood of the empirical distribution. When using neighborhoods based on relative entropy estimates, 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.

Key words and phrases. Bayesian inference, Mixture model, Model misspecification, Relative entropy, Robustness, Tempering.

1. INTRODUCTION

In many applications, the most natural models are idealizations that are known to provide only an approximation to the distribution of the observed data, due to small-scale contaminating effects that may be complicated and not completely understood. One might hope that any such lack of model fit, if sufficiently small, would not significantly impact inferences or decisions made based on the model. Often this does seem to be the case, but sometimes, unfortunately, the likelihood is strongly affected by perturbations to the distribution of the observed data, especially when the sample size is large.

As a result, standard Bayesian procedures are not generally robust to contamination or misspecification. In particular, this may lead to underestimation of uncertainty, since the posterior inexorably concentrates at a given rate—typically at minimal Kullback–Leibler points—regardless of whether or not it is concentrating on something that resembles the observed data distribution.

This issue was, perhaps, not as severe in the past, since on small datasets, statistical models automatically exhibit a certain amount of robustness to misspecification, because, in essence, there is insufficient power to discern small departures from the model. However, as datasets grow ever larger, even slight deficiencies in our models can become disruptive.

The problem is especially apparent in the context of Bayesian model averaging and flexible Bayesian models, such as nonparametric models. Under misspecification, model averaging tends to favor more complex models as the sample size grows, even when for all practical purposes the data are very well-described by a simpler model. Since it is usually unreasonable to expect real data to come exactly from a simple parametric model, we thus find ourselves in the awkward situation of eventually rejecting any such model.
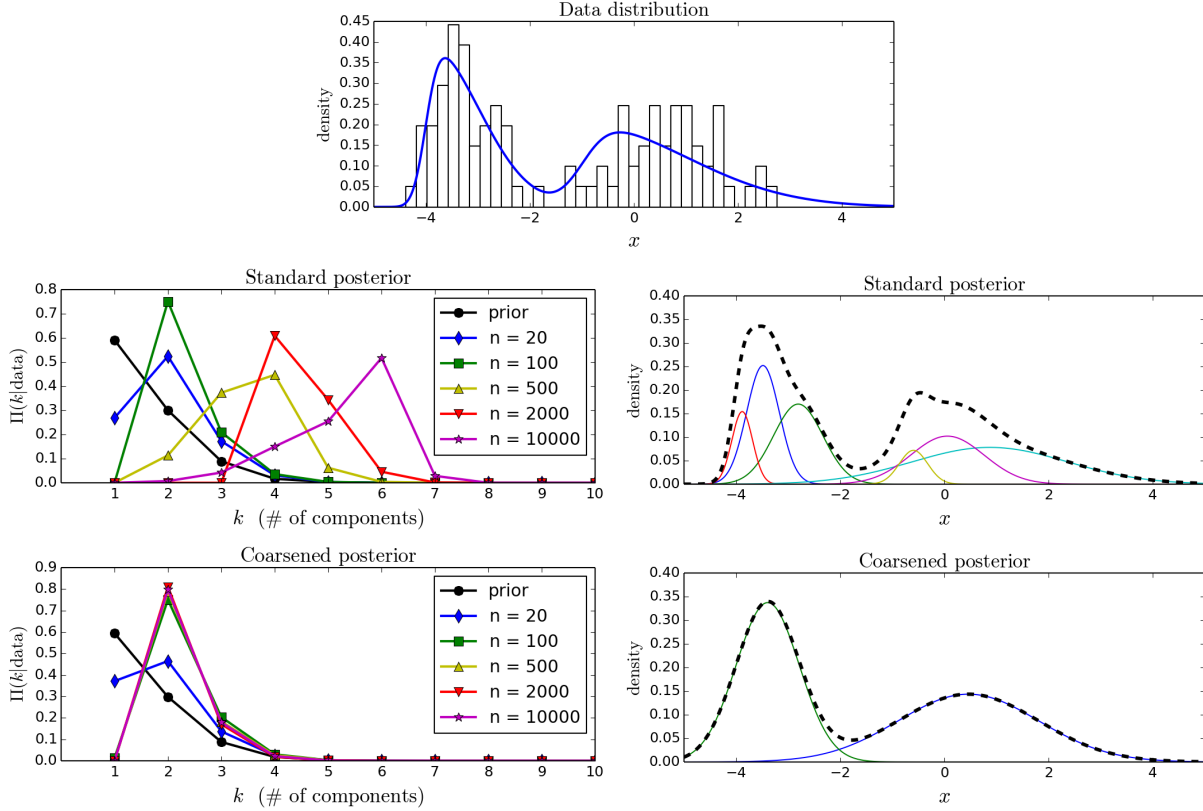
FIGURE 1. Gaussian mixture with a prior on the number of components k, applied to data from a two-component skew-normal mixture. Top: Density of the data distribution (blue line), and a histogram of $n = 100$ samples. Middle left: The posterior on k favors larger and larger values as n increases. Bottom left: The coarsened posterior on k stabilizes as n increases, favoring the true number of components, $k = 2$. Middle right: Mixture density (dotted black line) and components (solid colors) for a typical sample from the posterior when $n = 10^4$. Bottom right: Same for the coarsened posterior. See Section 7.3 for details.

At first, it may seem paradoxical that nonparametric models could suffer from misspecification issues, since after all, the point of a nonparametric model is that it can fit any distribution. It is true that if one is solely interested in fitting the data distribution—such as in density estimation or when the distribution is a nuisance parameter—then a nonparametric model will serve nicely, in principle. However, most nonparametric models still involve some parametric assumptions (such as Gaussian mixture components) and when the data do not accord well with these assumptions, the interpretability of parameters and latent variables (such as component parameters and cluster assignments) breaks down. Further, the computational burden of nonparametric models can grow rapidly with the sample size.

For example, suppose one is using a Gaussian mixture model with a prior on the number of components, but the data come from a mixture in which the components are not exactly Gaussian. In order to fit the data, the posterior will introduce more and more components as the amount of data increases, and the inferred components will not accurately reflect

the true components. To illustrate, suppose the data are from a two-component mixture of skew-normal distributions; see Figure 1. As shown in the figure, the posterior on the number of components favors larger and larger values as the sample size increases, and a typical sample from the posterior consists of many small components, obscuring the two large groups corresponding to the true components. Meanwhile, using the technique introduced in this paper, one can construct a "coarsened" posterior for the same model, under which the number of components does not continue to grow—see Figure 1—and as a result, one obtains a more macroscopic interpretation of the data as coming from two large groups. See Section 7.3 for details.

Ideally, one would model all aspects of the data generating process completely correctly, however, this is often impractical for a number of reasons. First, it may be unrealistic to expect to have sufficient insight into the data generating process to even write down an adequate model. Further, this can significantly increase the time and effort required to design the model, devise and implement reasonably efficient inference algorithms with reliable performance, and perhaps develop theoretical guarantees. Even after all such efforts, one may end up with a complex hierarchical model that is unlikely to be used in scientific applications, because scientists prefer statistical methods that have a clear and simple interpretation, and are not comfortable with drawing inferences from complicated models in a "black box" fashion.

In fact, in some cases, a simple model may actually be more appropriate than a more complex one, even when it does not exactly fit the observed data. For instance, when the underlying phenomenon of interest is well-described by a simple model, but there is a lack of fit due to contaminating artifacts, then choosing a more complex model can be viewed as a form of overfitting, and may damage the resulting inferences. Meanwhile, even when the underlying phenomenon is not well-described by a simple model, many times the purpose of a model is to provide a lens through which to understand the data, rather than just fitting it—such as when the model is being used as a tool for exploratory analysis—in which case it is essential to use interpretable models so that the parameters and latent variables provide insight into the questions of interest.

There have been advances in robustness to misspecification, with methods such as Gibbs posteriors (Jiang and Tanner, 2008), disparity-based posteriors (Hooker and Vidyashankar, 2014), partial posteriors (Doksum and Lo, 1990), nonparametric approaches (Rodríguez and Walker, 2014), neighborhood methods (Liu and Lindsay, 2009), and learning rate adjustment (Grünwald and van Ommen, 2014); see Section 4 for a discussion of previous work. However, despite growing recognition of the issue and efforts toward a solution, existing methods tend to be either limited in scope, computationally prohibitive, or lacking a clear justification.

These considerations lead to the following questions. Is it possible to draw coherent 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 explore 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 clearly incorrect when the model is misspecified—the approach we consider is, roughly speaking, to 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; see Section 3 for a detailed description.

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, automatically 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.

A particularly attractive case occurs when using neighborhoods based on relative entropy estimates, since then it turns out that the c-posterior can be approximated by simply raising the likelihood to a certain fractional power; see Section 3.1. Consequently, in this case 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 disadvantage of c-posteriors is that sometimes they are less concentrated than one would like—for instance, if it turns out that the amount of misspecification is less than expected.

An unexpected side benefit of our investigation of c-posteriors is that it reveals an interesting connection between Gibbs posteriors and approximate Bayesian computation (ABC), two areas of current research in Bayesian statistics. Roughly, a Gibbs posterior can be thought of as an asymptotic approximation to a particular ABC posterior; see Section 3.

The paper is organized as follows. First, to demonstrate the basic idea, in Section 2 we consider a c-posterior for the simplest possible toy example: Bernoulli trials. Then, in Section 3, we describe the c-posterior approach more generally. We discuss connections with previous work in Section 4, and in Section 5 we establish various theoretical properties of c-posteriors regarding asymptotics and robustness. In Section 6, we show how the method extends to time series and regression, and in Section 7, we apply the c-posterior approach to autoregressive models of unknown order, variable selection in linear regression, and mixture models with an unknown number of components. We close with a brief discussion of possible directions for future work.

2. TOY EXAMPLE: BERNOULLI TRIALS

For expository purposes, we first introduce the c-posterior in a toy example. 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 $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$.

The standard Bayesian approach is to define a prior probability for each hypothesis, say, $\Pi(H_0) = \Pi(H_1) = 1/2$, and define a prior density for θ in the case of H_1, say, $\theta|H_1 \sim$ Uniform$(0, 1)$. Inference then proceeds based on the posterior probabilities of the hypotheses, $\Pi(H_0|x_{1:n})$ and $\Pi(H_1|x_{1:n}) = 1 - \Pi(H_0|x_{1:n})$, 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, $\Pi(H_0|x_{1:n}) \xrightarrow{\text{a.s.}} \mathbb{1}(\theta = 1/2)$ as $n \to \infty$. (We use $\mathbb{1}(\cdot)$ to denote the indicator function: $\mathbb{1}(E) = 1$ if E is true, and $\mathbb{1}(E) = 0$ otherwise.)

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, as discussed in the introduction, and we seek a method that works well even with complex models.

Suppose it is known that any such 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 "true" data which do follow the model, and supposing that the observations $x_{1:n}$ are close to the true data in some distributional sense, but not necessarily equal to it. 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., $\hat{p}_x(1) = \bar{x}$ in this example.

Due to the corruption, it is inappropriate to condition on the true 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 $\Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x \| \hat{p}_X) < r\big)$. In other words, rather than conditioning on the data exactly, condition on a relative entropy neighborhood of the empirical distribution of the data.

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, or c-posterior, for inferences about H_0 and H_1:

$$(2.1) \qquad \Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x \| \hat{p}_X) < R\big).$$

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. A.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 2.1 can be computed exactly (see Section A.1 for details), however, in more complex cases, an approximation will be needed. In Section 3.1, we develop a general approximation which, when applied to this example, yields

$$(2.2) \qquad \Pi\big(\mathrm{H}_0 \,\big|\, D(\hat{p}_x \| \hat{p}_X) < R\big) \approx 1/\big(1 + 2^{\alpha_n} B(1 + \alpha_n \bar{x},\ 1 + \alpha_n(1 - \bar{x}))\big)$$

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

$$(2.3) \qquad \Pi\big(\mathrm{H}_0 \,\big|\, X_{1:n} = x_{1:n}\big) = 1/\big(1 + 2^n B(1 + n\bar{x},\ 1 + n(1 - \bar{x}))\big),$$

note that the only difference is that n has been replaced by α_n in the c-posterior.

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 2 (top and
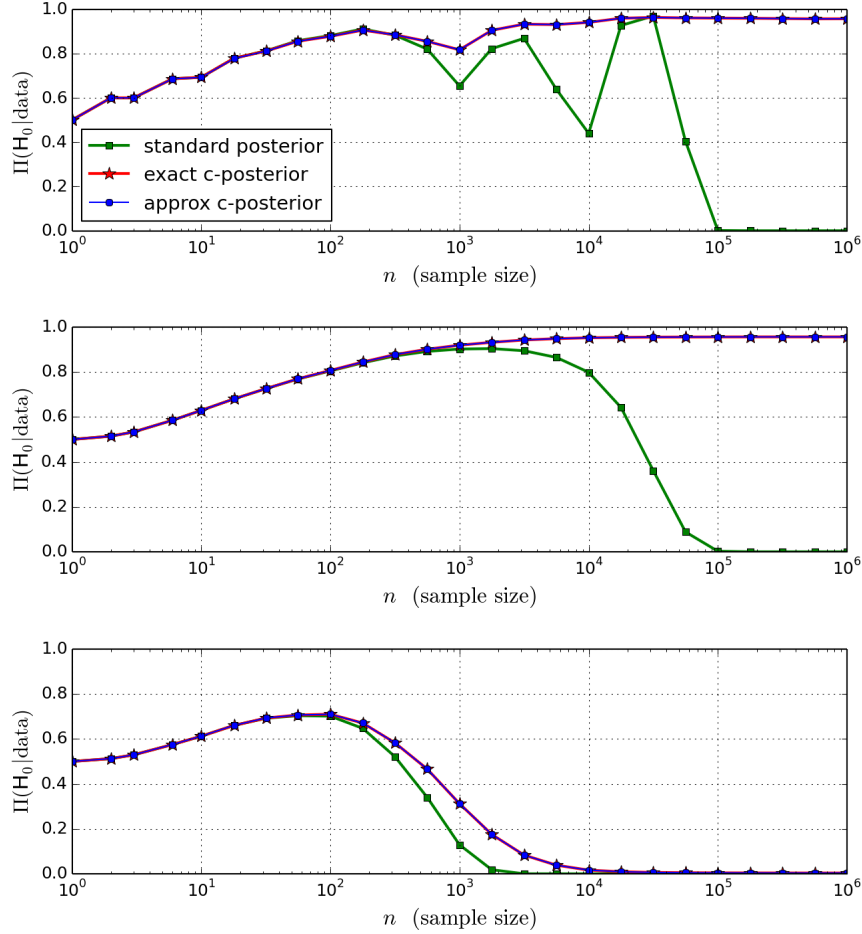
FIGURE 2. Bernoulli trials example. Top: Results from a single sequence $x_1, x_2, \ldots$ i.i.d. $\sim$ Bernoulli(0.51). Middle: Average over 1000 sequences $x_1, x_2, \ldots$ i.i.d. $\sim$ Bernoulli(0.51). Bottom: Same as middle, but with 0.56 instead of 0.51. In all three plots, the approximate c-posterior is indistinguishable from the exact c-posterior.

middle) shows the probability of H_0 under the standard posterior, the exact c-posterior, and the approximate c-posterior (Equations 2.3, 2.1, and 2.2, 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 2 (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 much data is observed.

3. Method

In this section, we describe the c-posterior approach more generally. For discussion of connections with previous work, see Section 4; in particular, our ideas have been influenced by Lindsay and Liu (2009) and Wilkinson (2013). For the time being, we assume i.i.d. data, but the approach generalizes, for example, to time-series (Section 6.1) and regression (Section 6.2).

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 merely be a useful fiction. (For context, note that in many scientific endeavors, one employs idealized models that capture the most important features of the phenomena of interest, without harboring any illusions that they completely describe every detail.) Now, suppose there are some unobserved *idealized data* $X_1, \ldots, X_n \in \mathcal{X}$ which 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 3.

If there were no corruption (i.e., contamination/misspecification), 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 clearly be incorrect. Of course, if one could easily model the 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, but this may be impractical, as discussed at length in the introduction.

An alternative approach is to condition on what is known—that 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 $\Pi(d\theta \mid X_{1:n} = x_{1:n})$, consider $\Pi\big(d\theta \mid d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < r\big)$. Since usually one will not have sufficient *a priori* knowledge to choose r, it makes sense to put a prior on it, say $R \sim H$, independently of θ and $X_{1:n}$. Generalizing further, take a sequence of functions d_n such that $d_n(X_{1:n}, x_{1:n}) \geq 0$ is some measure of the discrepancy between $X_{1:n}$ and $x_{1:n}$.

Definition 3.1. We refer to $\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$ as a *c-posterior*.

It is useful to note that one can write the c-posterior as

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

$$(3.1) \qquad\qquad = \Pi(d\theta) \int_{\mathcal{X}^n} G(d_n(x'_{1:n}, x_{1:n})) P_\theta^n(dx'_{1:n})$$

where $G(r) = \mathbb{P}(R > r)$. The intuitive interpretation is that, to use a rough analogy, this integral is like a convolution of P_θ^n (the distribution of $X_{1:n}$) with the "kernel" $G(d_n(X_{1:n}, x_{1:n}))$.
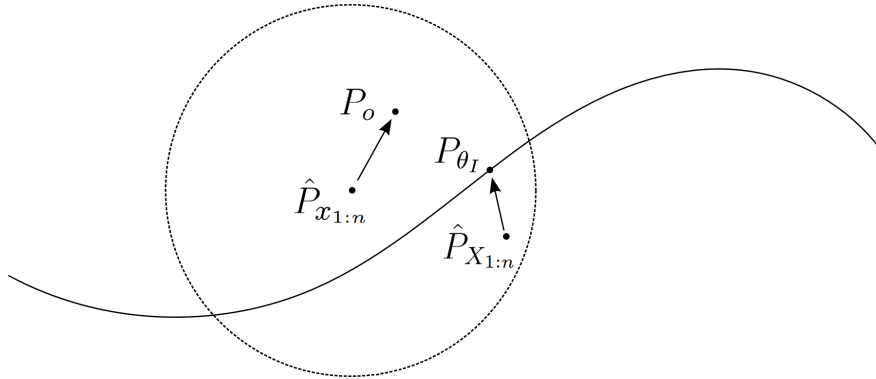
FIGURE 3. Notional schematic diagram of the idea behind the c-posterior. The ambient space is the set of probability distributions on $\mathcal{X}$, and the curve represents the subset of distributions in the parametrized family $\{P_\theta : \theta \in \Theta\}$. The idealized distribution P_{θ_I} is a point in this subset, and the empirical distribution $\hat{P}_{X_{1:n}}$ of the true 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 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.

Some readers will recognize the form of the c-posterior in connection with *approximate Bayesian computation* (ABC), where it arises due to the nature of the approximation (Marjoram et al., 2003). Needless to say, our motivation for using it is completely different than in ABC; see Section 4 for discussion.

While one could do inference for the c-posterior using algorithms similar to those used for ABC, this would be very slow, and would not exploit the fact that the likelihood is tractable. Instead, we develop approximations to the c-posterior that facilitate efficient inference. The crudest approach is based on the asymptotics of the c-posterior (see Section 5.1), which show that for large n,

$$(3.2) \qquad \Pi\big(d\theta \mid d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < R\big) \underset{\approx}{\propto} G(d(P_\theta, \hat{P}_{x_{1:n}}))\Pi(d\theta)$$

under mild regularity conditions; the intuition here is that when n is large, $\hat{P}_{X_{1:n}} \approx P_\theta$ with high probability. Thus, if $d(P_\theta, \hat{P}_{x_{1:n}})$ can be easily computed, Equation 3.2 can be used to perform approximate inference for the c-posterior by, for instance, using it as a target distribution in Metropolis–Hastings MCMC. If $R \sim \text{Exp}(\alpha)$, so that $G(r) = e^{-\alpha r}$, Equation 3.2 becomes $\exp(-\alpha d(P_\theta, \hat{P}_{x_{1:n}}))\Pi(d\theta)$, which some readers will recognize as a Gibbs posterior (Jiang and Tanner, 2008); see Section 4. More generally, similar approximations can be made when using a c-posterior based on $d_n(X_{1:n}, x_{1:n})$.

A major disadvantage of an asymptotic approximation as in Equation 3.2, however, is that it is only good when n is sufficiently large; in a sense, it ignores the randomness in $X_{1:n}$. For the case of relative entropy, we develop a much better approximation that is also applicable for smaller n; see Section 3.1 below.

There are many possible choices of statistical distance $d(\cdot, \cdot)$, and the robustness properties of the c-posterior depend on this choice; in Section 5.2, we show that as one would expect, the c-posterior is robust to changes in P_o that are small with respect to $d(\cdot, \cdot)$. We use the term

statistical distance very broadly, to mean any nonnegative function for assessing discrepancy that is meaningful for a given application. A few potential candidates would be Kolmogorov–Smirnov (in the univariate setting), Wasserstein, or a maximum mean discrepancy (Gretton et al., 2006). When P_θ and P_o admit density functions, it is also possible to accomodate 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 7), we focus on relative entropy (and variations thereof) as our choice of $d(\cdot, \cdot)$, since it works out exceptionally nicely in several respects. In particular, it turns out that in this case there is a trick that makes it unnecessary to explicitly compute $d_n(X_{1:n}, x_{1:n})$. We discuss this next.

3.1. Relative entropy c-posteriors.
In the case of relative entropy, there is an approximation to the c-posterior that improves upon Equation 3.2 in two respects: it is applicable for small n, and it is extremely easy to work with—in fact, it simply amounts to tempering the likelihood. As above, we assume i.i.d. data for the moment, and refer to Sections 6.1 and 6.2 for generalizations.

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. Define

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

and suppose $d_n(X_{1:n}, x_{1:n})$ is a consistent estimator of $D(p_o \| p_\theta)$. If $R \sim \mathrm{Exp}(\alpha)$, then asymptotically, the c-posterior based on $d_n(X_{1:n}, x_{1:n})$ is proportional to

(3.3)
$$\exp(-\alpha D(p_o \| p_\theta)) \Pi(d\theta) \propto \exp(\alpha \int p_o \log p_\theta) \Pi(d\theta)$$

$$\approx \exp \left(\alpha \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i) \right) \Pi(d\theta)$$

$$= \Pi(d\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\alpha/n}$$

under mild regularity conditions; see Section 5.1.2. When n is small relative to α, however, this is unsuitable as an approximation to the c-posterior, since in particular, if $\alpha/n > 1$ then this makes the likelihood more concentrated, rather than less.

Instead, we propose a better alternative, based on a central limit theorem approximation. The derivation is well-founded when the sample space $\mathcal{X}$ has finitely-many elements, and the extension to general $\mathcal{X}$, while heuristic, is intuitively sensible. 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. Assume $R \sim \mathrm{Exp}(\alpha)$. Then by Equation 3.1 and an approximation detailed in Section A.2,

$$\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) \propto \mathbb{E}\big(\exp(-\alpha D(\hat{p}_{x_{1:n}} \| \hat{p}_{X_{1:n}})) \mid \theta \big) \Pi(d\theta)$$

$$\approx (n\zeta_n/\alpha)^{\frac{|\mathcal{X}|-1}{2}} \exp(-n\zeta_n D(\hat{p}_{x_{1:n}} \| p_\theta)) \Pi(d\theta)$$

$$\propto \exp \big(\zeta_n \textstyle\sum_{i=1}^{n} \log p_\theta(x_i) \big) \Pi(d\theta)$$

where

(3.4)
$$\zeta_n = \frac{1/n}{1/n + 1/\alpha} = \frac{1}{1 + n/\alpha}.$$

The rough idea is that $\hat{p}_{X_{1:n}}$ is approximately multivariate normal with mean p_θ and precision of order n, and the exponentiated relative entropy approximates a normal density with mean $\hat{p}_{X_{1:n}}$ and precision of order α, so the expectation above behaves like the convolution of two normals, and the resulting precision is of order $1/(1/n + 1/\alpha) = n\zeta_n$. As demonstrated in Figure 2, this approximation can be quite accurate even when n is small, although more experimentation is needed to assess its accuracy in more general situations. Note that $\zeta_n \approx \alpha/n$ when $n \gg \alpha$, and $\zeta_n \approx 1$ when $\alpha \gg n$; thus, we refer to this as the small-sample correction to the asymptotic approximation.

This leads to the following useful approximation to the relative entropy c-posterior:

(3.5)
$$\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) \;\underset{\approx}{\propto}\; \Pi(d\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n}.$$

Note, in particular, that this 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 $\int p_o \log p_o$ term in $D(p_o\|p_\theta)$. Since this term is constant with respect to θ, it is absorbed into the constant of proportionality, allowing one to bypass this density estimation step, which would be both computationally expensive and statistically inefficient, especially in high dimensions.

Definition 3.2. Given $\zeta \in [0, 1]$, we refer to $\prod_{i=1}^{n} p_\theta(x_i)^\zeta$ as a *power likelihood*, and to the distribution proportional to $\Pi(d\theta) \prod_{i=1}^{n} p_\theta(x_i)^\zeta$ as a *power posterior*.

There are a number of other methods in which a power likelihood is employed (see Section 4), however, to our knowledge, the form of power we use (i.e., $\zeta = \zeta_n$), and its theoretical justification, are novel.

A useful interpretation of our power posterior is that it corresponds to adjusting the sample size from n to $n\zeta_n$. Thus, roughly speaking, by choosing a particular value of α, one makes the power posterior tolerant of all θ's for which a sample of size α from P_o could plausibly have come from P_θ. This idea is closely related to the model credibility index of Lindsay and Liu (2009). This suggests the interesting possibility of approximating the c-posterior by taking random subsets of size α, and combining the resulting posteriors in some way; this might have certain advantages in terms of computation and implementation, however, we do not explore it in this article.

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-occuring 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.

3.1.1. *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 $p_\theta(x) = \exp\big(\theta^{\mathrm{T}} s(x) - \kappa(\theta)\big)$, where $s(x) = (s_1(x), \ldots, s_k(x))^{\mathrm{T}}$ are the sufficient statistics, and suppose $\Pi(d\theta) = \pi_{\xi,\nu}(\theta)d\theta$ where $\pi_{\xi,\nu}(\theta) = \exp\big(\theta^{\mathrm{T}}\xi - \nu\kappa(\theta) - \psi(\xi, \nu)\big)$,

noting that this defines a conjugate family. Then the power posterior is proportional to

$$(3.6) \qquad \pi_{\xi,\nu}(\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n} \propto \exp\left(\theta^{\mathrm{T}}\big(\xi + \zeta_n \textstyle\sum_i s(x_i)\big) - (\nu + n\zeta_n)\kappa(\theta) \right) \propto \pi_{\xi_n,\nu_n}(\theta),$$

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,

$$\int_{\Theta} \pi_{\xi,\nu}(\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n} d\theta = \exp\left(\psi(\xi_n, \nu_n) - \psi(\xi, \nu) \right),$$

which can be used to compute robustified Bayes factors or a robust posterior on models, in the context of model inference. This is robust to small perturbations to P_o (in the sense of relative entropy), whereas the usual model inference/selection procedures can be very sensitive to such perturbations, for large n; see Section 5.2 for details.

In Section 2, we used this approach in the toy example involving Bernoulli trials. In Section 7.1, we apply it to perform robust inference for the order of an autoregressive model.

3.1.2. *MCMC on the power posterior.* Often, it is desirable to place conditionally-conjugate priors on the parameters of an exponential family—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 3.6—the full conditionals belong to the conjugate family, making them easy to sample from. In Section 7.2, we use Gibbs sampling for 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 7.3, we use Metropolis–Hastings for inference in mixtures with a prior on the number of components, with the power posterior.

By a stroke of luck, the mixing time for MCMC with the power posterior will often be better than with the standard posterior, since raising the likelihood to a fractional power (i.e., 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 time, a well-known example being MC^3 (Geyer, 1991).

Thus, 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 noted. 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 Gibbs sampling from 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.

4. Connections 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)—indeed, ABC provided part of our inspiration for considering this form of posterior. However, there are some crucial distinctions to note. First, the motivation here is completely different than with 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 is easily computed, which makes inference much more computationally efficient. Another major difference is that in ABC, the coarsened posterior is viewed as an undesirable side effect of the approximate nature of ABC, while from our perspective, it is precisely the object of interest—in other words, for us it is an asset, not a liability.

The c-posterior can also be viewed as conditioning on partial information, a technique which is often used to improve robustness (Doksum and Lo, 1990; Pettitt, 1983; Hoff, 2007; Dunson and Taylor, 2005); 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 very 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, for n sufficiently large, the c-posterior is approximately proportional to $\exp(-\alpha d(P_\theta, \hat{P}_{x_{1:n}}))\Pi(d\theta)$, 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 fact that Gibbs posteriors can be interpreted as an approximation to a coherent Bayesian procedure (the c-posterior) 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 started to be 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 Grünwald (2011, 2012). Nonetheless, 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.

Conceptually speaking, the existing methods that are perhaps most similar to the idea of the c-posterior are goodness-of-fit tests that assess whether the data distribution is within a neighborhood of the model space (Rudas et al., 1994; Goutis and Robert, 1998; Dette and Munk, 2003; Liu and Lindsay, 2009), however, the methods used previously are very different from ours. Closely related to such work is the model credibility index of Lindsay and Liu (2009), a concept which has heavily influenced our thinking in the development of the c-posterior.

5. THEORY

In this section, we establish the asymptotic form of c-posteriors (Section 5.1) and their robustness properties (Section 5.2). 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 consider the following model:

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

where H is a distribution on $[0, \infty)$. Note that we use (bold) $\boldsymbol{\theta}$ for the random variable, and θ for particular values. Define

$$G(r) = \mathbb{P}(R > r).$$

Now, 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.

5.1. **Asymptotic form of the c-posterior.** The c-posterior takes a simple form as $n \to \infty$, under mild regularity conditions. We prove a general convergence theorem for c-posteriors (Theorem 5.3) and then apply it to c-posteriors derived from relative entropy (Corollary 5.4) and weakly-continuous distances (Corollary 5.5).

5.1.1. *Convergence theorem.* The following basic lemma captures the underlying principle at work in establishing both the asymptotic form of the c-posterior (Theorem 5.3) as well as its robustness (Theorem 5.6).

Lemma 5.1. *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*

$$\mathbb{E}(W \mid U_n < V) \xrightarrow[n\to\infty]{} \mathbb{E}(W \mid U < V).$$

All proofs for this section have been placed in Section A.3. The following condition is necessary to avoid certain pathologies; it is always satisfied, for instance, 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 $[0, \infty)$.

Condition 5.2. *Assume $\mathbb{P}(d(P_{\boldsymbol{\theta}}, P_o) = R) = 0$ and $\mathbb{P}(d(P_{\boldsymbol{\theta}}, P_o) < R) > 0$.*

We use $\Rightarrow$ to denote convergence with respect to the weak topology.

Theorem 5.3. *If $d_n(X_{1:n}, x_{1:n}) \xrightarrow{\text{a.s.}} d(P_{\boldsymbol{\theta}}, P_o)$ as $n \to \infty$ and Condition 5.2 is satisfied, then*

$$(5.1) \qquad \Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) \xRightarrow[n \to \infty]{} \Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big) \propto G\big(d(P_\theta, P_o)\big)\Pi(d\theta),$$

and in fact,

$$(5.2) \quad \mathbb{E}\big(h(\boldsymbol{\theta}) \mid d_n(X_{1:n}, x_{1:n}) < R\big) \xrightarrow[n \to \infty]{} \mathbb{E}\big(h(\boldsymbol{\theta}) \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big) = \frac{\mathbb{E}h(\boldsymbol{\theta})G\big(d(P_{\boldsymbol{\theta}}, P_o)\big)}{\mathbb{E}G\big(d(P_{\boldsymbol{\theta}}, P_o)\big)}$$

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

As noted earlier, a case of particular interest arises when $R \sim \mathrm{Exp}(\alpha)$, 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 5.3. This is asymptotically equivalent to $\exp(-\alpha d(P_{\boldsymbol{\theta}}, \hat{P}_{x_{1:n}}))\Pi(d\theta)$, provided that $d(P_{\boldsymbol{\theta}}, \hat{P}_{x_{1:n}}) \xrightarrow{\text{a.s.}} d(P_{\boldsymbol{\theta}}, P_o)$, and as discussed in Section 4, this is precisely the form of a Gibbs posterior; thus, a Gibbs posterior can be interpreted as a large-sample approximation to a c-posterior. It is also worth noting that if $R = r_0$ with probability 1 for some $r_0 > 0$, then $G(r) = \mathbb{1}(r < r_0)$, and by Theorem 5.3 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 r_0 neighborhood of P_o and reverts to the prior inside.

5.1.2. *Application to relative entropy.* Suppose P_o has density p_o and P_θ has density p_θ for each $\theta \in \Theta$.

Corollary 5.4. *Suppose $d_n(X_{1:n}, x_{1:n})$ 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 5.2 is satisfied, then Equations 5.1 and 5.2 hold.*

This establishes the asymptotic form of the relative entropy c-posterior as claimed in Equation 3.3.

5.1.3. *Application to weakly-continuous distances.* Recall that $\hat{P}_{x_{1:n}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ denotes the empirical distribution of $x_{1:n}$.

Corollary 5.5. *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 5.2 is satisfied, then Equations 5.1 and 5.2 hold when $d_n(X_{1:n}, x_{1:n}) = d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}})$.*

5.2. **Robustness properties.** Here, we show that 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 inference (Section 5.2.1), while c-posteriors are robust to small changes in P_o (Section 5.2.2). To see roughly why the standard posterior is not robust, note that

$$\Pi(d\theta \mid X_{1:n} = x_{1:n}) \propto \exp\Big(\sum_{i=1}^{n} \log p_\theta(x_i)\Big)\Pi(d\theta) \approx \exp\big(n\textstyle\int p_o \log p_\theta\big)\Pi(d\theta)$$

$$\propto \exp(-nD(p_o\|p_\theta))\Pi(d\theta),$$

assuming the densities p_θ and p_o exist. 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 $\exp(-\alpha D(p_o\|p_\theta))\Pi(d\theta)$, suggesting that the c-posterior should remain stable in the limit as $n \to \infty$; below we make this precise, for a general choice of $d(\cdot, \cdot)$.

5.2.1. *Lack of robustness of model inference under the standard posterior.* 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

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

as $n \to \infty$, where $\theta_k^n = \mathrm{argmax}_{\theta \in \Theta_k} p(x_{1:n}|\theta)$ is the maximum likelihood estimator for model k, $\theta_k^* = \mathrm{argmin}_{\theta \in \Theta_k} D(p_o\|p_\theta)$ is the minimal Kullback–Leibler (KL) point within model k, and $H(\theta; p_o) = -\int p_o(\nabla_\theta^2 \log p_\theta)$. 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

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

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 inference with the standard posterior is not robust.

5.2.2. *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 asymptotically, the outcome of an inference procedure be continuous as a function of P_o, with respect to some topology (the weak topology being a standard choice) (Huber, 2004). From this perspective, the lack of robustness of the standard posterior can be thought of as a lack of continuity with respect to P_o, asymptotically.

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 will be robust to perturbations to P_o, provided that $d(\cdot, \cdot)$ is chosen appropriately. In the terminology of Section 3, if the observed data distribution P_o is close to the ideal data 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 5.6. *If $P_1, P_2, \ldots \in \mathcal{M}$ such that $d(P_\theta, P_m) \xrightarrow[m \to \infty]{} d(P_\theta, P_o)$ for Π-almost all $\theta \in \Theta$, and Condition 5.2 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, $\Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_m) < R\big) \Longrightarrow \Pi\big(d\theta \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big)$.

In particular, Theorem 5.6 implies that the c-posterior is robust in the context of model inference, since taking $h(\theta) = \mathbb{1}(\theta \in \Theta_k)$, we have

$$\Pi\big(\Theta_k \mid d(P_{\boldsymbol{\theta}}, P_m) < R\big) \longrightarrow \Pi\big(\Theta_k \mid d(P_{\boldsymbol{\theta}}, P_o) < R\big)$$

as $m \to \infty$, under the assumptions of the theorem.

6. EXTENSIONS

6.1. **Time-series c-posterior based on relative entropy rate.** Suppose the sequence of observed data $(x_1, \ldots, x_n)$ is a partial sample from a stationary and ergodic process with distribution P_o, and suppose the model $\{P_\theta : \theta \in \Theta\}$ consists of stationary finite-order Markov processes. Assume that for some sigma-finite measure μ on $\mathcal{X}$, for all $n \in \{1, 2, \ldots\}$ and all $\theta \in \Theta$, the finite-dimensional distributions have densities $p_o(x_1, \ldots, x_n)$ and $p_\theta(x_1, \ldots, x_n)$ with respect to the product measure μ^n, and assume $\mathbb{E}_{P_o}|\log p_o(X_{1:n})| < \infty$ and $\mathbb{E}_{P_o}|\log p_\theta(X_{1:n})| < \infty$.

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

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

Suppose $d_n(X_{1:n}, x_{1:n})$ is an a.s.-consistent estimator of $\mathcal{D}(P_o \| P_\theta)$ when $(X_1, X_2, \ldots) \sim P_\theta$ and $(x_1, x_2, \ldots) \sim P_o$, and consider the c-posterior $\Pi\big(d\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$, with $R \sim \text{Exp}(\alpha)$. Then by Lemma 5.1, 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). Combining this with the small-sample correction (applied heuristically in this setting) suggests the approximation

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

$$\propto \Pi(d\theta) \prod_{i=1}^{n} p_\theta(x_i|x_1,\ldots,x_{i-1})^{\zeta_n}.$$

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 7.1, we apply this to perform robust inference for the order of an autoregressive model.

6.2. **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 allow for contamination/misspecification, let us suppose that $Y_i|x_i$ is drawn from the model $p_\theta(y|x)$ for $i = 1,\ldots,n$, and 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 $(x_1, y_1),\ldots,(x_n, y_n)$ 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 5.1, the resulting c-posterior converges to a nice asymptotic form:

$$\Pi\big(d\theta \mid d_n(Y_{1:n}, y_{1:n}|x_{1:n}) < R\big) \implies \Pi\big(d\theta \mid D_\theta < R\big) \propto \exp(-\alpha D_\theta)\Pi(d\theta)$$

$$\propto \exp\Big(\alpha \int p_o(x, y) \log p_\theta(y|x) \, dx \, dy\Big)\Pi(d\theta)$$

if we take $R \sim \text{Exp}(\alpha)$ as usual. To obtain an approximation that is applicable for smaller n as well, we apply the same small-sample correction as before, replacing α by $n\zeta_n$. Combining this with an empirical approximation to the integral suggests using

$$\Pi\big(d\theta \mid d_n(Y_{1:n}, y_{1:n}|x_{1:n}) < R\big) \underset{\sim}{\propto} \exp\big(\zeta_n \textstyle\sum_i \log p_\theta(y_i|x_i)\big)\Pi(d\theta)$$

$$= \Pi(d\theta) \prod_{i=1}^{n} p_\theta(y_i|x_i)^{\zeta_n}.$$

Consequently, once again, we arrive at a power posterior approximation to the c-posterior, allowing us to bypass the computation of $d_n(\cdot, \cdot|\cdot)$. In Section 7.2, we apply this to perform robust variable selection in linear regression.

7. APPLICATIONS

7.1. **Autoregressive models of unknown order.** We illustrate using 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 noise/shocks, but also to misspecification of the structure of the model, such as time-varying noise. This serves as a nice 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.

We apply the time-series c-posterior developed in Section 6.1 to obtain robustness to perturbations that are small in the sense of relative entropy rate. Since $\beta | k$ has been given a conjugate prior, then as described in Section 3.1.1, we can analytically compute the resulting marginal power likelihood,

$$p_c(x_{1:n}|k) := \int_{\mathbb{R}^k} p(x_{1:n}|\theta, k)^{\zeta_n} \pi(\theta|k) d\theta$$

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

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

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 $v_i = \sum_{t=1}^{n} x_t x_{t-i}/\sigma^2$, by straightforward calculation. This, in turn, can be used to compute a robustified posterior on the model order k, defined as $\pi_c(k|x_{1:n}) \propto p_c(x_{1:n}|k)\pi(k)$. This is expected to be robust to departures from the $AR(k)$ model that require more than α samples to distinguish, and thus, it will 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:

$$(7.1) \qquad\qquad x_t = \sum_{\ell=1}^{4} \theta_\ell x_{t-\ell} + \varepsilon_t + \tfrac{1}{2} \sin t$$

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, this can be used to make a principled choice of α; see Section 7.2. Here, we assess sensitivity to the choice of α, by considering the c-posterior on k as α varies, when $n = 10^4$, with an improper uniform prior on k; see Figure 4 (upper right). 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 4 (lower left).
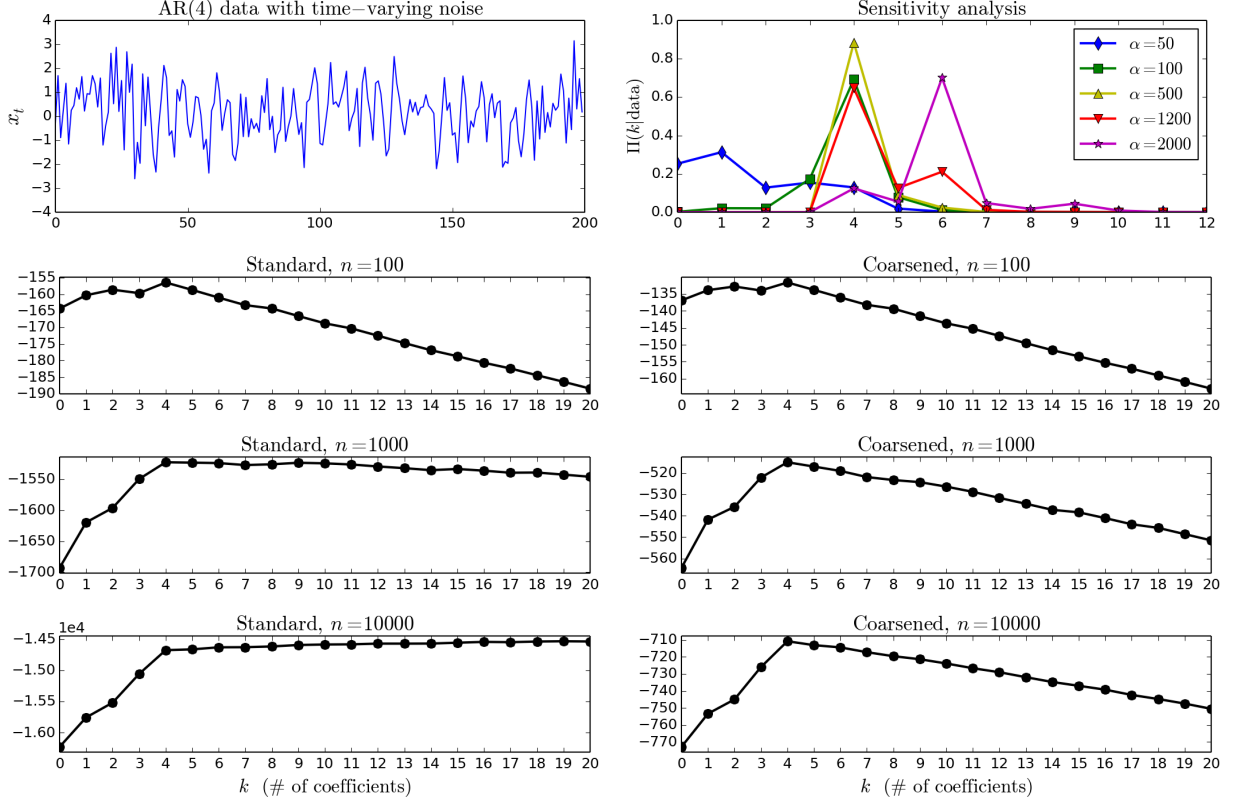
FIGURE 4. Autoregression example. Upper left: Data sampled from the process in Equation 7.1. Upper right: Sensitivity analysis, displaying the c-posterior on k as α varies, when $n = 10^4$. Lower left: Log marginal likelihood of AR(k) model for $k = 0, 1, \ldots, 20$, on increasing amounts of data from this process. Lower right: Log of coarsened marginal likelihood for the same model, on the same data.

Meanwhile, the c-posterior stabilizes to a distribution on k favoring $k = 4$; see Figure 4 (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 type of picture is typical for the log marginal likelihood when comparing models of increasing complexity. As discussed in Section 5.2, the log marginal likelihood automatically penalizes more complex models, via the term $-\frac{1}{2}t_k \log n$ where t_k is the dimension of the parameter space, 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 only increases proportionally to $\log n$, 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, continuing to increase 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.

7.2. Variable selection in linear regression.

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

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

Models of this type are often used to infer which covariates $x_{i1}, \dots, x_{ip}$ 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 $\beta^{\mathsf{T}} x_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip}$. 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.

As described in Section 6.2, in the regression setting, the c-posterior based on conditional relative entropy can be approximated by the power posterior obtained by raising the likelihood to $\zeta_n = 1/(1 + n/\alpha)$, as before. If we first integrate W out of the model, the resulting power posterior is

$$\pi_c(\beta, \lambda|y) \propto \pi(\beta, \lambda) p(y|\beta, \lambda)^{\zeta_n}.$$

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:

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

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

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

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 $\mathcal{N}(M, L^{-1})$.

7.2.1. Simulation example.

To demonstrate empirically, first consider a simulated example where the mean of the observed data is a slightly nonlinear function of a single covariate, plus a constant offset:

$$(7.2) \qquad\qquad y_i = \beta_{01} + \beta_{02}(x_{i2} + \tfrac{1}{16} x_{i2}^2) + \varepsilon_i$$

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 accomodate 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) which 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 5 (top) 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 the coarsening parameter α, recall that the c-posterior is obtained by conditioning on the (estimated) conditional relative entropy being less than R, where $R \sim \text{Exp}(\alpha)$. 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 5 (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 7.2 is $k = 2$ (β_{01} and β_{02}).

Figure 5 (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, under misspecification, 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 amount of concentration; however, the price to be paid is that this can cause the c-posterior to be considerably more diffuse than necessary, such as in the case of β_2.

7.2.2. *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,
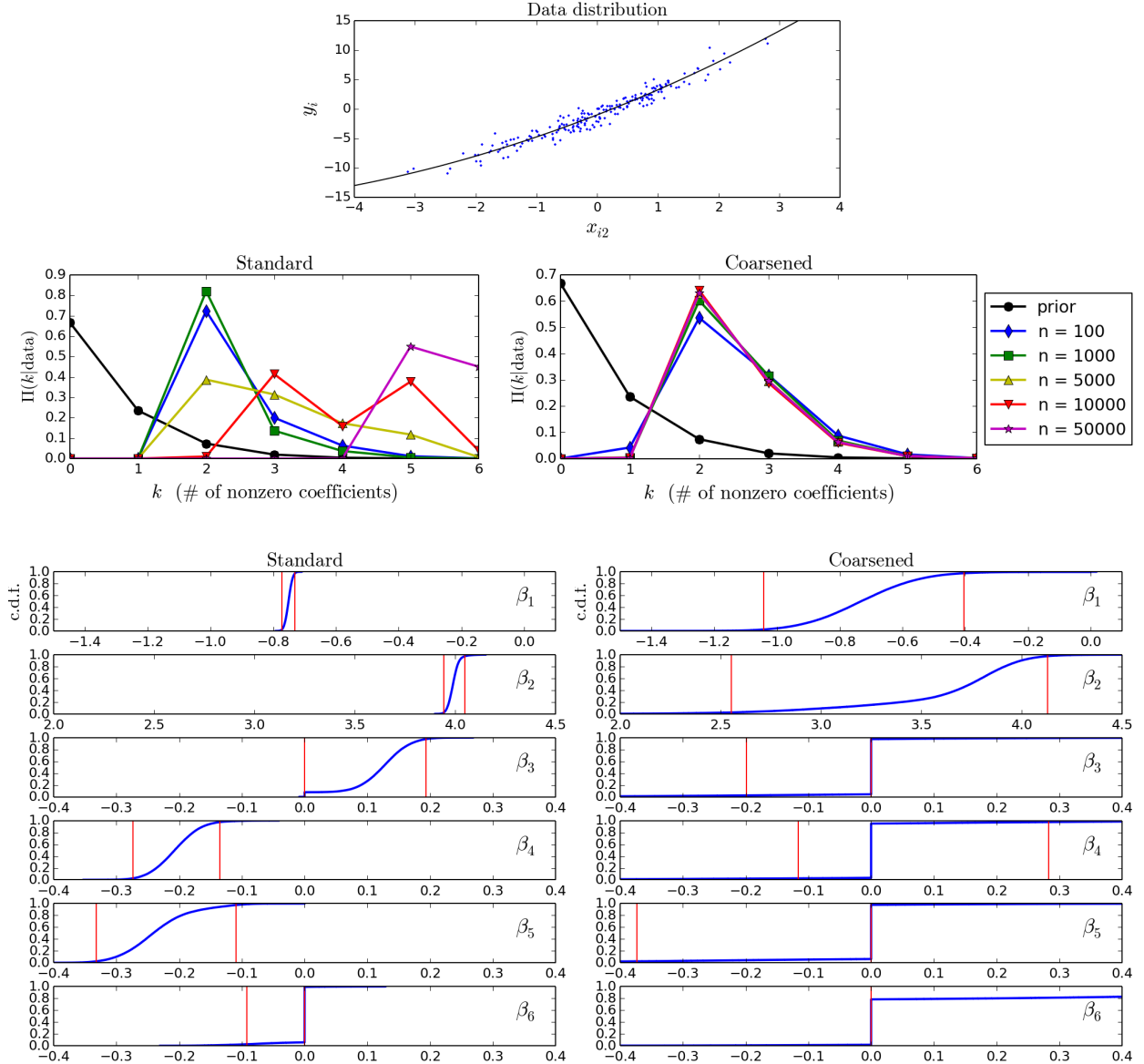
FIGURE 5. Variable selection with simulated data. Top: Scatterplot of the target variable y_i versus x_{i2}, as well as the mean function (black line). Middle left: The posterior on the number of nonzero coefficients k favors larger values as n increases. Middle right: The c-posterior favors the "true" 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.

2009). Using a subset of the CPP data, 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
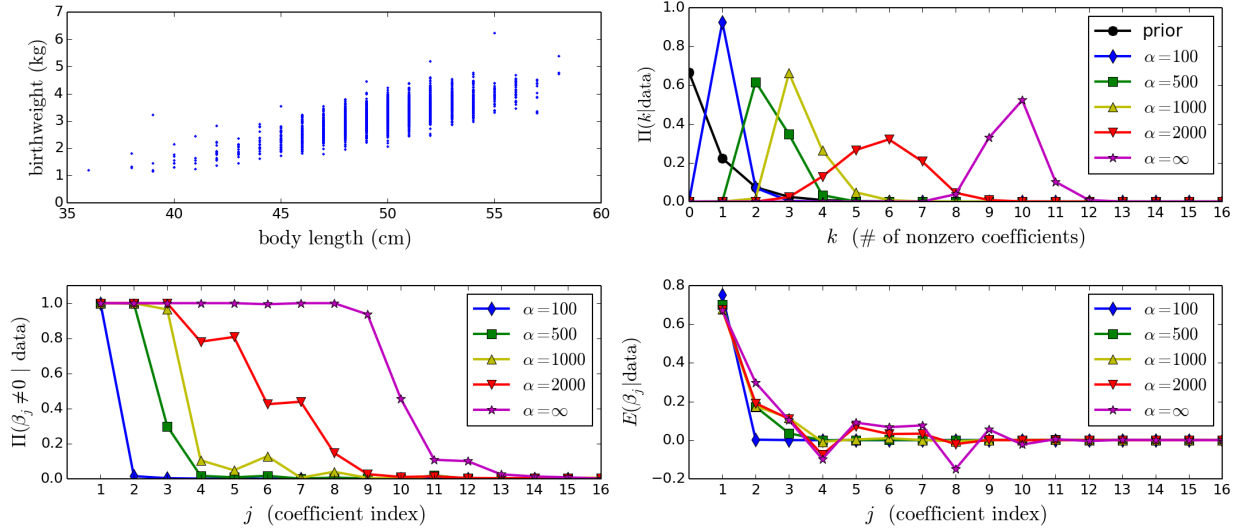
FIGURE 6. Variable selection for modeling birthweight. Upper left: Scatter-plot of birthweight (the target variable) versus body length at birth. Upper right: The standard posterior includes several more nonzero coefficients than the c-posteriors. Lower left: Posterior probability of inclusion for each 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.)

same model 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 $\lambda \approx 2.5$ to 3, and thus, by the formula $\alpha \approx 2\sigma^2/\delta^2 = 2/(\lambda\delta^2)$ derived above, 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 6, upper right), and the posterior probability of inclusion (Figure 6, 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 6 (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 $\delta_{\mathrm{kg}} = 0.01$ 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, and their values, becomes less related to practically significant associations and more related to the fact that the model is compensating for its limitations.

7.3. Mixture models with a prior on the number of components. Consider a finite mixture model:

$$X_1, \ldots, X_n | k, w, \varphi \ \text{i.i.d.} \sim \sum_{i=1}^{k} w_i f_{\varphi_i}(x),$$

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 ($f_\varphi : \varphi \in \Phi$), resulting in 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

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

The usual approaches to inference in mixtures are based on latent variables indicating which component each datapoint comes from, but unfortunately, that does not seem to work here. There are, nonetheless, a few possible approaches to doing inference, for instance, Antoniano-Villalobos and Walker (2013) developed an auxiliary variable technique for posteriors of this form, and it would also be possible to use reversible jump MCMC (Green, 1995). To keep things as simple as possible, however, we assume an upper bound on k, say $k \leq m$, and reparameterize 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 $w_i = g(v_i) / \sum_{j=1}^{m} g(v_j)$ and $g(v) = \max\{v - c, 0\}$, so that $w_i = 0$ if $v_i \leq c$. Letting $v_1, \ldots, v_m \sim \mathrm{Gamma}(a, b)$ i.i.d., conditioned 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 k (that is, the number of nonzero weights) is $\pi(k) \propto \mathrm{Binomial}(k | m, p) \mathbb{1}(k > 0)$ where $p = \mathbb{P}(v_i > c)$.

7.3.1. *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 the two-component mixture $\frac{1}{2}\mathcal{SN}(-4, 1, 5) + \frac{1}{2}\mathcal{SN}(-1, 2, 5)$, where $\mathcal{SN}(\xi, s, a)$ is the skew-normal distribution (Azzalini and Capitanio, 1999) with location ξ, scale s, and shape a; see Figure 1 (top). For the model parameters, we assume an upper bound of $m = 10$ 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 in the preceding paragraph, with $a = 1/m$, $b = 1$, and c such that $p = \mathbb{P}(v_i > c) = 1/m$.

For the c-posterior, we set $\zeta_n = 1/(1+n/\alpha)$ (following Equation 3.4), 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. The top panel in Figure 1 illustrates that based on 100 samples, one can visually determine that there are two large groups of roughly equal size, and can roughly determine their location and scale, but cannot determine their precise form—in particular, one cannot tell that they are not actually Gaussian.

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 both the standard and coarsened posteriors. Each sweep consisted of MH moves on each (μ_i, λ_i) and v_i separately. Figure 1 (left) shows the average of the approximated posteriors on k over these five datasets. For typical samples from the standard and coarsened posteriors when $n = 10000$, Figure 1 (right) shows the mixture density $\sum_{i=1}^{m} w_i f_{\mu_i, \lambda_i}(x)$, and the individual (weighted) components $w_i f_{\mu_i, \lambda_i}(x)$.

As expected, 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 obtain an adequate fit to the data. Meanwhile, in accordance with our visual intuition that, based on 100 samples, there appear to be two large groups, the $\alpha = 100$ c-posterior on k shows strong support for two components, no matter how large n becomes.

The typical sample from the standard posterior provides a better fit to the distribution of the data, however, it has several more than two components, obscuring the fact that there are two large groups. On the other hand, the typical sample from the c-posterior does not fit the data distribution as well, but it clearly indicates that there are two groups, and provides an interpretable representation of their locations and scales.

7.3.2. *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 $n = 82$. 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 7. 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 the coarsening parameter α, one can explore the data at varying levels of precision, allowing for greater or smaller departures from normality.

We use the same model as above, 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
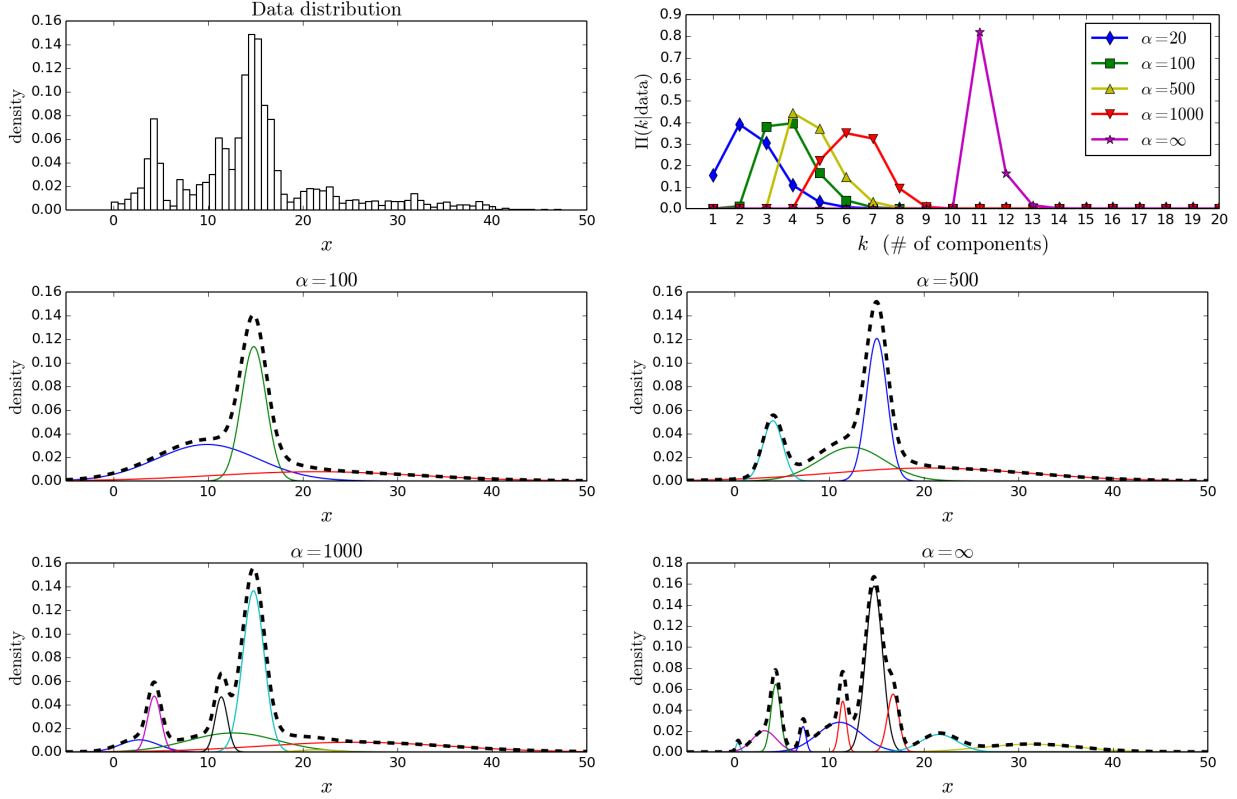
FIGURE 7. Gaussian mixture with a prior on the number of components k, applied to the Shapley galaxy data. Top left: Histogram of the data, in units of 1,000 km/s, excluding a small amount of data extending in a tail up to 80,000 km/s. Top right: C-posterior on k for a range of α values; $\alpha = \infty$ 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.

($\alpha = \infty$), 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 7, 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.

8. Discussion

The c-posterior approach seems promising as a general method of robust Bayesian inference, being both computationally efficient and conceptually well-grounded. There are a number of directions that would be interesting to pursue in future work. Further investigation of the small-sample correction is needed—in particular, assessing its accuracy, and justifying its use in non-discrete cases, if possible. 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 nice if there was a way of inferring the coarsening parameter α from data, somehow; this would enable one not only to perform robust inference, but also to infer the amount of misspecification, and ideally, to achieve statistical efficiency when the model is correct. It would be beneficial, but perhaps difficult, if precise guarantees could be provided regarding frequentist coverage properties of the c-posterior, under misspecification. Finally, the scope of application of this way of thinking seems potentially larger than Bayesian inference; it might be interesting to explore adaptations to frequentist procedures.

Appendix A. Proofs

A.1. **C-posterior for the toy example in Section 2.** Letting $Z = \mathbb{1}(D(\hat{p}_x||\hat{p}_X) < R)$, by Bayes' theorem we have that for $h \in \{H_0, H_1\}$,

$$\Pi\big(h|Z=1\big) \underset{h}{\propto} \mathbb{P}(Z=1|h)\Pi(h) \overset{(a)}{\underset{h}{\propto}} \mathbb{P}(Z=1|h)$$

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

where (a) is since $\Pi(h) = 1/2$, and (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 $S|H_0 \sim \text{Binomial}(n, 1/2)$ and $S|H_1 \sim \text{BetaBinomial}(n, 1, 1) = \text{Uniform}\{0, 1, \ldots, n\}$. To derive the approximation in Equation 2.1, we use Equation A.1 below:

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

where $\alpha_n = 1/(1/n + 1/\alpha)$, $p_\theta(x) = \text{Bernoulli}(x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x \in \{0,1\}$, and c is a constant that does not depend on θ or h. If $h = H_1$, then this yields

$$\mathbb{P}(Z=1|H_1) = \mathbb{E}\Big(\mathbb{E}\big(\exp(-\alpha D(\hat{p}_x||\hat{p}_X))\,\big|\,\theta, H_1\big)\,\Big|\,H_1\Big) \approx \mathbb{E}\Big(c\prod_{i=1}^n p_\theta(x_i)^{\alpha_n/n}\,\Big|\,H_1\Big)$$

$$= c\int_0^1 \theta^{\alpha_n \bar{x}}(1-\theta)^{\alpha_n(1-\bar{x})}\,d\theta = cB\big(1 + \alpha_n\bar{x},\, 1 + \alpha_n(1-\bar{x})\big).$$

If $h = H_0$, then $\theta = 1/2$ with probability 1, so $\mathbb{P}(Z=1|H_0) \approx c(1/2^{\alpha_n/n})^n = c/2^{\alpha_n}$. Thus, $\Pi(H_0|Z=1) = \mathbb{P}(Z=1|H_0)/(\mathbb{P}(Z=1|H_0) + \mathbb{P}(Z=1|H_1)) \approx 1/(1 + 2^{\alpha_n}B(1 + \alpha_n\bar{x},\, 1 + \alpha_n(1-\bar{x}))$.

A.2. **Justification of the small-sample correction when $|\mathcal{X}| < \infty$.** 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 $X_1, \dots, X_n$ i.i.d. $\sim s$ and $\hat{\boldsymbol{s}}_j = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i = j)$ for $j = 1, \dots, k$, then for $p \in \Delta_k$ near s,

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

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 $x' = (x_1, \dots, x_{d-1})$. First, for $q \in \Delta_k$ near p,

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

by Propositions A.1 and A.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 $C = C(s')$,

$$\mathbb{E} \exp(-\alpha D(p\|\hat{\boldsymbol{s}})) \approx \mathbb{E} \exp(-\alpha D(p\|\boldsymbol{q})) \mathbb{1}(\boldsymbol{q} \in \Delta_k)$$

$$\overset{(a)}{\approx} \mathbb{E} \exp\left(-\frac{\alpha}{2}(p' - \boldsymbol{q}')^{\mathsf{T}} C^{-1}(p' - \boldsymbol{q}') \right)$$

$$= (2\pi)^{\frac{k-1}{2}} |C/\alpha|^{1/2} \int \mathcal{N}(p'|q', C/\alpha) \mathcal{N}(q'|s', C/n) dq'$$

$$\overset{(b)}{=} (2\pi)^{\frac{k-1}{2}} |C/\alpha|^{1/2} \mathcal{N}(p'|s', (1/\alpha + 1/n)C)$$

$$= \left(\frac{1/\alpha}{1/\alpha + 1/n} \right)^{\frac{k-1}{2}} \exp\left(-\tfrac{1}{2}(1/\alpha + 1/n)^{-1}(p' - s')^{\mathsf{T}} C^{-1}(p' - s') \right)$$

$$\overset{(c)}{\approx} (n\zeta_n/\alpha)^{\frac{k-1}{2}} \exp(-n\zeta_n D(p\|s)),$$

where (a) is by Equation A.2 along with the approximation $C(\boldsymbol{q}') \approx C(s')$, (b) uses the convolution formula for independent normals, and (c) is again by Equation A.2. This yields Equation A.1.

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 A.1. *For $p, q \in \Delta_k$, $D(p\|q) = \tfrac{1}{2}\chi^2(p,q) + o(\|p - q\|^2)$ 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,

$$f(a) = f(b) + f'(b)(a - b) + \tfrac{1}{2}f''(b)(a - b)^2 + o(|a - b|^2)$$

$$= (a - b) + \frac{1}{2}\frac{(a - b)^2}{b} + o(|a - b|^2)$$

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 $p \to q$. $\qquad \square$

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 A.2. *For any $p, q \in \Delta_k$, $\chi^2(p, q) = (p' - q')^\mathsf{T} C^{-1}(p' - q')$ 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, $C^{-1} = (\mathrm{diag}(q') - q' q'^\mathsf{T})^{-1} = \mathrm{diag}(q')^{-1} + (1/q_k)\mathbf{1}\mathbf{1}^\mathsf{T}$ where $\mathbf{1} = (1, \ldots, 1)^\mathsf{T}$, hence

$$(p' - q')^\mathsf{T} C^{-1}(p' - q') = \sum_{i=1}^{k-1} \frac{(p_i - q_i)^2}{q_i} + \frac{\left(\sum_{i=1}^{k-1}(p_i - q_i)\right)^2}{q_k}$$

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

A.3. Proofs of results in Theory section.

Proof of Lemma 5.1. 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

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

$\square$

Proof of Theorem 5.3. We apply Lemma 5.1 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 5.1,

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

since $V \perp U, W$ by construction. This establishes Equation 5.2, and since in particular this holds for any bounded continuous h, Equation 5.1 follows. $\square$

Proof of Corollary 5.5. 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 5.3 applies. $\square$

Proof of Theorem 5.6. Apply Lemma 5.1 with $U = d(P_{\boldsymbol\theta}, P_o)$, $U_m = d(P_{\boldsymbol\theta}, P_m)$, $V = R$, and $W = h(\boldsymbol\theta)$. $\square$

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.

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.

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

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

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

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.

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.

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.

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

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. Safe Learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *COLT*, volume 19, pages 397–419, 2011.

Grünwald, P. The safe Bayesian: learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*, volume 7568, pages 169–183. Springer, 2012.

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

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

Hooker, G. and Vidyashankar, A. N. Bayesian model robustness via disparities. *TEST*, 23 (3):556–584, 2014.

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.

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.

Rodríguez, C. E. and Walker, S. G. Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Statistics and Computing*, 24(1):35–49, 2014.

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.

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.

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.