Published in final edited form as:
J Am Stat Assoc. 2019 ; 114(527): 1113–1125. doi:10.1080/01621459.2018.1469995.

Robust Bayesian inference via coarsening

Jeffrey W. Miller,
Department of Biostatistics, Harvard University

David B. Dunson
Department of Statistical Science, Duke University

Abstract

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

Keywords

Model error; model misspecification; power likelihood; relative entropy; tempering; clustering

1 Introduction

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

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

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

It works out particularly well to use relative entropy (i.e., Kullback–Leibler divergence), since in this case the c-posterior can be approximated by the "power posterior" obtained by simply raising the likelihood to a certain fractional power. Consequently, one can usually do approximate inference using standard algorithms with no additional computational burden—in fact, the mixing time of Markov chain Monte Carlo (MCMC) samplers will typically be improved, since the likelihood is tempered. Further, when using exponential families and conjugate priors, one can even obtain analytical expressions for quantities such as a robustified marginal likelihood.

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

The paper is organized as follows. Section 2 introduces the coarsening approach and considers the case of relative entropy coarsening in detail. Section 3 uses a toy Bernoulli example to illustrate coarsening in the simplest possible setting, as well as to assess the accuracy of the power posterior approximation. Section 4 introduces a technique for choosing an appropriate amount of coarsening in a data-driven way. In Section 5, we demonstrate coarsening for mixture models and clustering, to obtain robustness to the form of the component distributions. We apply this to perform robust clustering of cells in flow cytometry datasets containing tens of thousands of multivariate data points. In Section 6, we demonstrate coarsening on autoregressive models of unknown order, performing inference for the model complexity in a way that is robust to perturbations. Section 7 discusses several frequently asked questions, and the supplementary material contains theoretical results, previous work, further discussion, and additional details.

2 Method

For now, we assume an i.i.d. setting, but the approach generalizes to time series and regression (see Section 6 and Supplement S6). Suppose we have a model $\{P_\theta : \theta \in \Theta\}$ along with a prior $\prod$ on Θ, and suppose there is a point $\theta_I \in \Theta$ representing the parameters of the *idealized distribution* of the data. The interpretation is that θ_I is the true state of nature about which one is interested in making inferences. Suppose there are some unobserved *idealized data* $X_1, \ldots, X_n \in \mathcal{X}$ that are i.i.d. from P_{θ_I}, and the *observed data* $x_1, \ldots, x_n \in \mathcal{X}$ are a perturbed version of $X_1, \ldots, X_n$ in the sense that $d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < r$ for some discrepancy $d(\cdot, \cdot)$ and some $r > 0$, where $\hat{P}_{x_{1:n}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ denotes the empirical distribution of $x_{1:n} = (x_1, \ldots, x_n)$. Suppose $x_1, \ldots, x_n$ behave like i.i.d. samples from some P_o, which we view as a perturbation of P_{θ_I}. For intuition, consider the diagram in Figure 1.

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

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

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

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

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

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

In Supplement S3.1, we derive the form of the c-posterior as $n \to \infty$. Meanwhile, in Supplement S3.2, we show that under certain conditions, when n is fixed and the distribution of R converges to 0, the c-posterior converges to the standard posterior. In Supplement S3.3, we show that the c-posterior is robust to changes in P_o that are small with respect to the chosen discrepancy $d(\cdot, \cdot)$. There are different types of robustness that may be desired, and the type of robustness exhibited by the c-posterior can be customized through the choice of $d(\cdot, \cdot)$. A few potential candidates for $d(\cdot, \cdot)$ would be Kolmogorov–Smirnov (in the univariate setting), Wasserstein, or a maximum mean discrepancy (Gretton et al., 2006). When P_θ and P_o have densities with respect to a common measure, it is also possible to accommodate discrepancies between densities such as relative entropy, Hellinger distance, and various divergences—even though they may be undefined for empirical distributions—by choosing $d_n(X_{1:n}, x_{1:n})$ to be a consistent estimator of $d(P_\theta, P_o)$.

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

2.1 Relative entropy c-posteriors

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

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

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

$$\pi(\theta \mid d_n(X_{1:n}, x_{1:n}) < R) \underset{\sim}{\propto} \pi(\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n}, \qquad (2.2)$$

where $\underset{\sim}{\propto}$ means “approximately proportional to”, i.e., the distribution on the left is approximately equal to the distribution proportional to the expression on the right, and

$$\zeta_n = \frac{1/n}{1/n + 1/\alpha} = \frac{\alpha}{\alpha + n}. \qquad (2.3)$$

The approximation in Equation 2.2 is good when either $n \gg \alpha$ or $n \ll \alpha$ (Corollary S3.4, Theorem S3.6), under mild conditions. Empirically we find that the approximation can be quite accurate (see Figure 2). It makes intuitive sense that the approximation would be good in both the large-sample ($n \gg \alpha$) and small-sample ($n \ll \alpha$) regimes, when one considers the convolution representation in Equation 2.1. Also, note that $\zeta_n \approx \alpha/n$ when $n \gg \alpha$, whereas $\zeta_n \approx 1$ when $n \ll \alpha$, and ζ_n smoothly interpolates between these two regimes. For the motivation behind the particular form of the power ζ_n, see Supplement S5. Section 4 introduces a technique for choosing α in a data-driven way.

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

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

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

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

2.1.1 Power posterior with conjugate priors—

When using exponential families with conjugate priors, one can often obtain analytical expressions for integrals with respect to the power posterior. Suppose $p_\theta(x) = \exp(\theta^\mathrm{T} s(x) - \kappa(\theta))$, where $s(x) = (s_1(x), \ldots, s_k(x))^\mathrm{T}$ are the sufficient statistics, and suppose $\pi(\theta) = \pi_{\xi,\nu}(\theta)$ where $\pi_{\xi,\nu}(\theta) = \exp(\theta^\mathrm{T}\xi - \nu\kappa(\theta) - \psi(\xi, \nu))$, noting that this defines a conjugate family. Then the power posterior is proportional to

$$\pi_{\xi,\nu}(\theta)\prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n} \propto \exp\left(\theta^\mathrm{T}\left(\xi + \zeta_n\sum_i s(x_i)\right) - (\nu + n\zeta_n)\kappa(\theta)\right) \propto \pi_{\xi_n,\nu_n}(\theta), \qquad (2.4)$$

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

For most conjugate families used in practice, simple analytical expressions are available for the log-normalization constant $\psi(\xi, \nu)$ as well as for many integrals with respect to $\pi_{\xi,\nu}(\theta)$.

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

2.1.2 MCMC on the power posterior—

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

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

3 Toy example: Perturbed Bernoulli trials

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

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

A natural choice of discrepancy is the relative entropy between the empirical distributions of $x_{1:n}$ and $X_{1:n}$, $D(\hat{p}_x \parallel \hat{p}_X) = \sum_{i=0}^{1} \hat{p}_x(i) \log(\hat{p}_x(i) \big/ \hat{p}_X(i))$, where $\hat{p}_x(1) = \bar{x}$ and $\hat{p}_x(0) = 1 - \bar{x}$ in this example. This leads us to consider the following coarsened posterior for inferences about H_0 and H_1:

$$\prod(H_0 \mid D(\hat{p}_x \parallel \hat{p}_X) < R), \tag{3.1}$$

where $R \sim \text{Exp}(\alpha)$. How should we choose α? If we have no *a priori* knowledge of the size of perturbation to expect, then we can use the calibration curve technique in Section 4. Otherwise, in this example, we can interpret the neighborhood size r in terms of Euclidean distance via the chi-squared approximation to relative entropy, $D(p \parallel q) \approx \frac{1}{2}\chi^2(p, q)$ (see Prop. S5.1). In particular, when $\bar{X} \approx 1 \big/ 2$ we have $D(\hat{p}_x \parallel \hat{p}_X) \approx 2 \mid \bar{x} - \bar{X} \mid^2$. Thus, if we expect that the perturbation will shift the sample mean by no more than ε or so when $H_0 : \theta = 1/2$ is true, then it makes sense to choose α so that $\mathbb{E}R \approx 2\varepsilon^2$. Since $\mathbb{E}R = 1 \big/ \alpha$, this suggests using $\alpha = 1/(2\varepsilon^2)$.

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

$$\prod\left(H_0 \mid D(\hat{p}_x \parallel \hat{p}_X) < R\right) \approx 1 \bigg/ \left(1 + 2^{n\zeta_n} B(1 + n\zeta_n \bar{x}, 1 + n\zeta_n(1 - \bar{x}))\right) \tag{3.2}$$

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

$$\prod(H_0 \mid X_{1:n} = x_{1:n}) = 1 \Big/ (1 + 2^n B(1 + n\bar{x}, 1 + n(1 - \bar{x}))), \tag{3.3}$$

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

To illustrate numerically, suppose we would like to be robust to perturbations affecting $\bar{x}$ by roughly $\varepsilon = 0.02$ when H_0 is true. As described above, this corresponds to $\alpha = 1/(2 \cdot 0.02^2) = 1250$. Now, suppose that in reality H_0 is indeed true (i.e., the true distribution is $P_{\theta_I} = \text{Bernoulli}(\theta_I)$ where $\theta_I = 0.5$), and the data are perturbed in such a way that $x_1, \ldots, x_n$ behave like i.i.d. samples from Bernoulli(θ^o) where $\theta^o = 0.51$ (i.e., the observed data distribution is $P_o = \text{Bernoulli}(\theta^o)$). Figure 2 (top left) shows the probability of H_0 under the standard posterior, the exact c-posterior, and the approximate c-posterior (Equations 3.3, 3.1, and 3.2, respectively), for increasing values of the sample size n.

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

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

4 Calibration curve technique for choosing α

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

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

The three calibration curves in Figure 2 illustrate common patterns. Case (i): When there is no perturbation from H_0, the best fit is obtained with very low complexity at the terminus $\alpha = \infty$. This suggests choosing $\alpha = \infty$, which would make the c-posterior concentrate at the true value in this case. Case (ii): When there is a small perturbation from H_0 ($\theta^o = 0.51$), the fit increases dramatically at first while maintaining low complexity, then the curve reaches a cusp at $\alpha \approx 2500$ and levels off, with more modest increases in fit at the cost of greater complexity. This suggests choosing $\alpha \approx 2500$. The curve sits near the cusp for a large range of α values from around 1200 to 4000, e.g., the blue dot indicates $\alpha = 1250$, our *a priori* choice. Any value of α in this range yields similar results (e.g., see Figure 2 bottom right compared to top left). Case (iii): When the distance from H_0 is very large ($\theta^o = 0.75$), there is no cusp in the curve, and a good fit can only be obtained at higher complexity. This curve suggests choosing $\alpha = \infty$, in which case the c-posterior would concentrate at H_1. This makes sense since the distance from H_0 is so large that explaining it by a perturbation is implausible. Thus, the calibration curve can help decide how much coarsening is needed, if any.

5 Mixture models and clustering

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

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

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

Here, we explore two algorithms: (a) a conditional coarsening algorithm and (b) an importance sampling algorithm for the power posterior. The conditional coarsening algorithm scales well, is easy to implement, and yields results similar to (but not exactly the same as) the power posterior. It is identical to the standard data augmentation algorithm for mixtures, except that the updates to w and φ use a power likelihood.

Algorithm 5.1 (Conditional coarsening for mixture models).

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

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

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

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

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

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

5.1 Simulation example: Perturbed mixture of Gaussians

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

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

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

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

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

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

Figure 3(a,b) shows the calibration curves for the $k_0 = 2$ and $k_0 = 4$ examples, with $n = 10^4$ data points. In both examples, there is a clear cusp at a point of good fit and low complexity. In the $k_0 = 2$ example, the curve is near the cusp when α is around 800 to 1000, and the tip is at $\alpha \approx 800$; thus, we choose $\alpha = 800$ in this example. In the $k_0 = 4$ example, a wide range of α values from 800 to 2000 are near the cusp, with the tip at $\alpha \approx 2000$; thus, we pick $\alpha =$ 2000 in this case.

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

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

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

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

5.2 Application: Robust clustering for flow cytometry

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

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

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

We explore a coarsening approach to robust clustering for flow cytometry data, using a multivariate Gaussian mixture model. For the model parameters, in the same notation as at the beginning of Section 5, we use $K = 20$, $\gamma_1 = \cdots = \gamma_K = 0.5/K$, and component parameter priors $\mu_i \sim \mathcal{N}(m, L^{-1})$ and $\Lambda_i \sim \text{Wishart}(V, \nu)$ independently, where the component densities are $f_{\mu_i, \Lambda_i}(x) = \mathcal{N}(x \mid \mu_i, \Lambda_i^{-1})$ for $x \in \mathbb{R}^d$. We set the hyperparameters in a data-dependent way: given input data $x_1, \ldots, x_n \in \mathbb{R}^d$, we choose prior mean $m = \frac{1}{n}\sum_{j=1}^n x_j$, prior precision matrix $L = \left(\frac{1}{n}\sum_{j=1}^n (x_j - m)(x_j - m)^{\mathrm{T}}\right)^{-1}$, degrees of freedom $\nu = d$, and scale matrix $V = L/\nu$. Algorithm 5.1 is implemented by defining $\varphi_i = (\mu_i, \Lambda_i)$ and using power-likelihood Gibbs updates to μ_i and Λ_i for step 3 of the algorithm:

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

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

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

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

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

For a range of α values, for each training dataset, we run Algorithm 5.1 and at each iteration we compute $F(\mathscr{A}, \mathscr{B})$ with $\mathscr{A}$ as the manual ground truth and $\mathscr{B}$ as the partition induced by z^*. In each dataset, a small fraction of cells were not labeled by the human expert; these unlabeled cells are included when running the algorithm, and excluded when computing the F-measure. Figure 5 shows the average F-measure for each of these runs, excluding burn-in. Averaging over the six training datasets, the best performance is obtained at $\alpha = 200$; thus, we set $\alpha = 200$ to evaluate performance on the test datasets.

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

6 Autoregressive models of unknown order

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

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

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

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

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

$$x_t = \sum_{\ell=1}^{4} \theta_\ell x_{t-\ell} + \varepsilon_t + \frac{1}{2}\sin t \qquad (6.1)$$

where $\theta = (1/4, 1/4, -1/4, 1/4)$, $\varepsilon_t \overset{\mathrm{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, 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, we set $\sigma_0^2 = 1$, and we use a Geometric(0.1) prior on k (i.e., $\pi(k) = 0.9^k 0.1$ for $k \in \{0, 1, 2, \ldots\}$).

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

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

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

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

7 Discussion

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

Concentration versus calibration

It is important to note that, unlike the standard posterior, the c-posterior does not concentrate as $n \to \infty$. This is appropriate and desirable, since if there is a perturbation then some uncertainty always remains about the true (idealized) distribution, no matter how much data are observed. Thus, in practice, the best one can do is appropriately quantify uncertainty about the true distribution, which is precisely what the c-posterior is designed to do, by allowing for perturbations.

However, if too much coarsening is applied (e.g., if α is too small), then (1) model complexity estimates under the c-posterior will tend to be biased downward, and (2) posterior credible sets will be overly large, since the c-posterior will not be sufficiently concentrated. These are the main disadvantages of using a c-posterior.

Thus, it is necessary to appropriately calibrate the amount of coarsening to match the size of the perturbation. If the model is exactly correct, then any amount of coarsening would be disadvantageous. However, in practice, the model is almost always wrong, so some amount of coarsening would probably be beneficial in most applications.

Bias

For any given level of model complexity, parameter estimates under the c-posterior have very similar bias to standard posterior estimates, as long as the prior is not overly strong. For instance, in regression problems, the c-posterior mean of the regression function tends to be very close to the standard posterior mean function — the c-posterior is just less concentrated about this function.

Meanwhile, the bias of model complexity estimates can be considerably improved under the c-posterior. This is illustrated by the mixture model example and the autoregression example, in which the bias of the posterior mean of k is massively reduced by using a c-posterior rather than the standard posterior.

Not equivalent to renormalized tempering or overdispersion

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

Measurement error

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

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

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

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

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

distributional perturbations even as $n \to \infty$; essentially, it needs to be a nonparametric model for the empirical distribution $\hat{P}_{x_{1:n}}$ given $\hat{P}_{X_{1:n}}$ (see Figure 1). But this seems just as computationally burdensome as using a nonparametric model for P_o given P_{θ_I}, and then modeling $x_1, \ldots, x_n$ as i.i.d. from P_o.

Separation of the amount of coarsening from the choice of prior

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

The technical reason is that the likelihood overwhelms the prior as the sample size increases. Thus, no matter what prior is selected, a perturbation involving the likelihood will require the same amount of coarsening. A robust prior provides robustness to the choice of prior, but not robustness to the choice of likelihood. For example, in variable selection linear regression models, we have observed that using a mixture of g priors (a leading example of a robust prior) is still just as sensitive to perturbations as using a more informative prior. Coarsening addresses the problem by dealing with the likelihood directly.

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

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

Overfitting

In the coarsening approach, we can choose the discrepancy function and α in addition to the likelihood and prior, and one may wonder whether this could lead to overfitting. Usually, having more choices makes a method more flexible. However, coarsening can be viewed as a form of regularization, so in fact, it reduces overfitting and makes the results less sensitive to modeling choices. To make a very rough analogy to penalized regression, the choices involved in coarsening (namely, the discrepancy and α) are analogous to the form of regularization penalty (e.g., ridge, lasso, elastic net) and the regularization parameter, λ. The point of coarsening is that it reduces sensitivity to model assumptions, by allowing for perturbations of the model. The discrepancy and α simply specify in what way you want to reduce sensitivity (i.e., what kind of perturbations you want robustness against).

8 Conclusion

The c-posterior approach seems promising as a general method of robust Bayesian inference. There are several directions that would be interesting to pursue in future work. Instead of using a single α for the entire likelihood, one could potentially use different amounts of coarsening on different parts of the likelihood, since some parts may be more misspecified than others. Further investigation of the accuracy of the power posterior approximation is needed, both theoretically and empirically. Additionally, it would be beneficial if precise guarantees could be provided regarding frequentist coverage properties of the c-posterior when there is a perturbation. Finally, it would be interesting to explore coarsening in frequentist procedures, since the scope of application is not limited to Bayesian inference.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

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

References

Aghaeepour N Nikolic R Hoos HH Brinkman RR Rapid cell population identification in flow cytometry data Cytometry Part A 2011 79 1 6–13

Aghaeepour N Finak G Hoos H Mosmann TR Brinkman R Gottardo R Scheuermann RH FlowCAP Consortium DREAM Consortium Critical assessment of automated flow cytometry data analysis techniques Nature Methods 2013 10 3 228–238 91 others [PubMed: 23396282]

Antoniano-Villalobos I Walker SG Bayesian nonparametric inference for the power likelihood Journal of Computational and Graphical Statistics 2013 22 4 801–813

Brinkman RR Gasparetto M Lee S-JJ Ribickas AJ Perkins J Janssen W Smiley R Smith C High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease Biology of Blood and Marrow Transplantation 2007 13 6 691–700 [PubMed: 17531779]

Carroll, RJ, Ruppert, D, Stefanski, LA, Crainiceanu, CM. Measurement Error in Nonlinear Models: A Modern Perspective. CRC press; 2006.

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

Green PJ Reversible jump Markov chain Monte Carlo computation and Bayesian model determination Biometrika 1995 82 4 711–732

Gretton A Borgwardt KM Rasch M Schölkopf B Smola AJ A kernel method for the two-sample-problem Advances in Neural Information Processing Systems 2006 19 513–520

Ishwaran H Zarepour M Dirichlet prior sieves in finite normal mixtures Statistica Sinica 2002 12 941–963

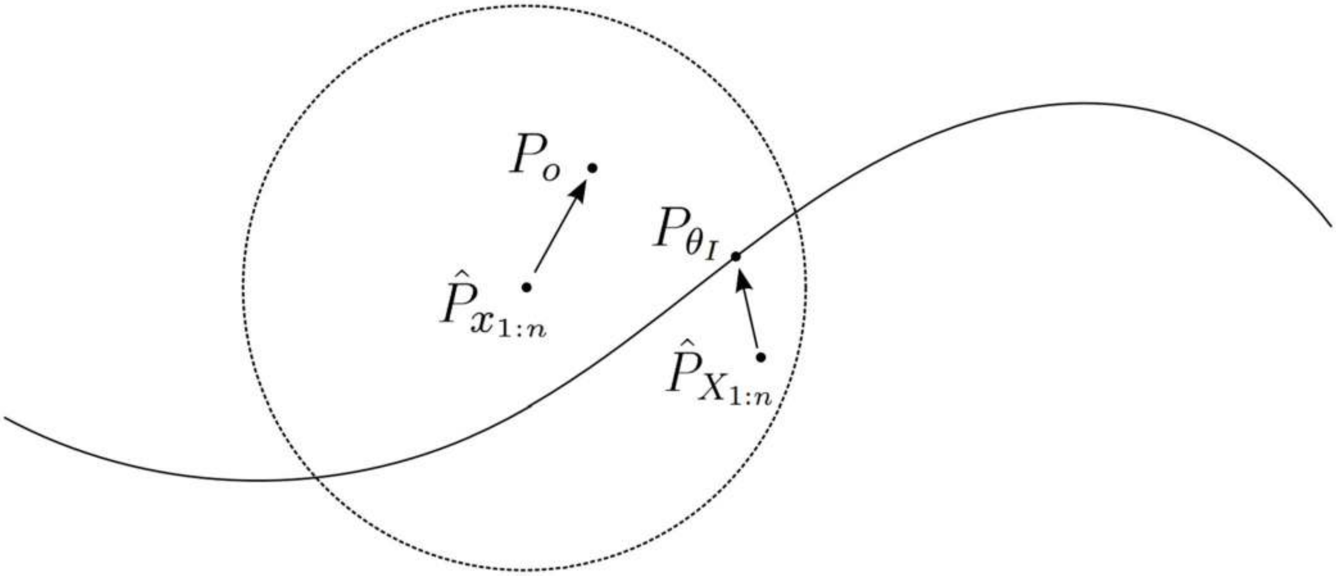


Figure 1.

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

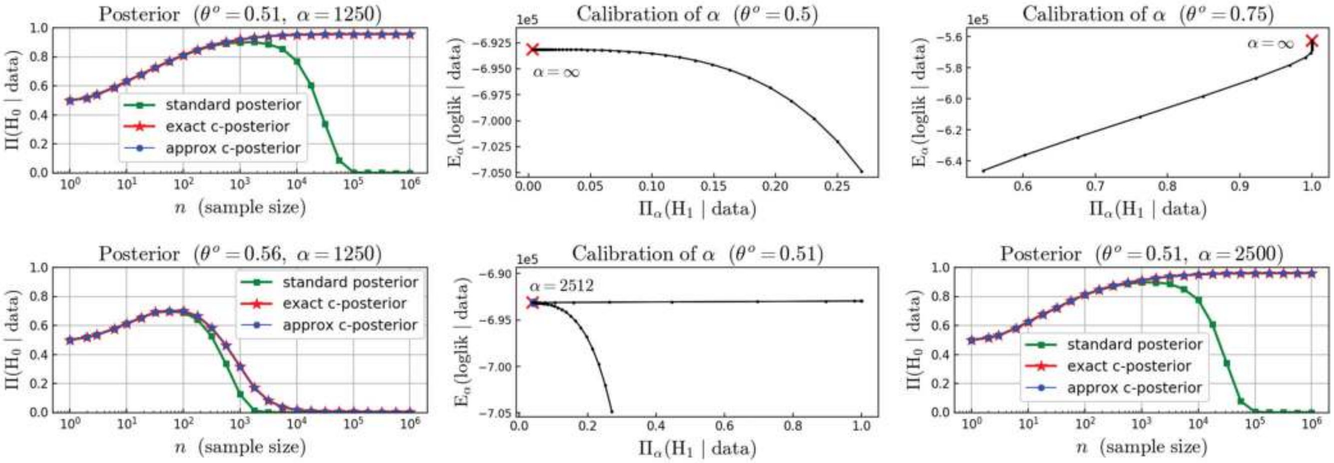
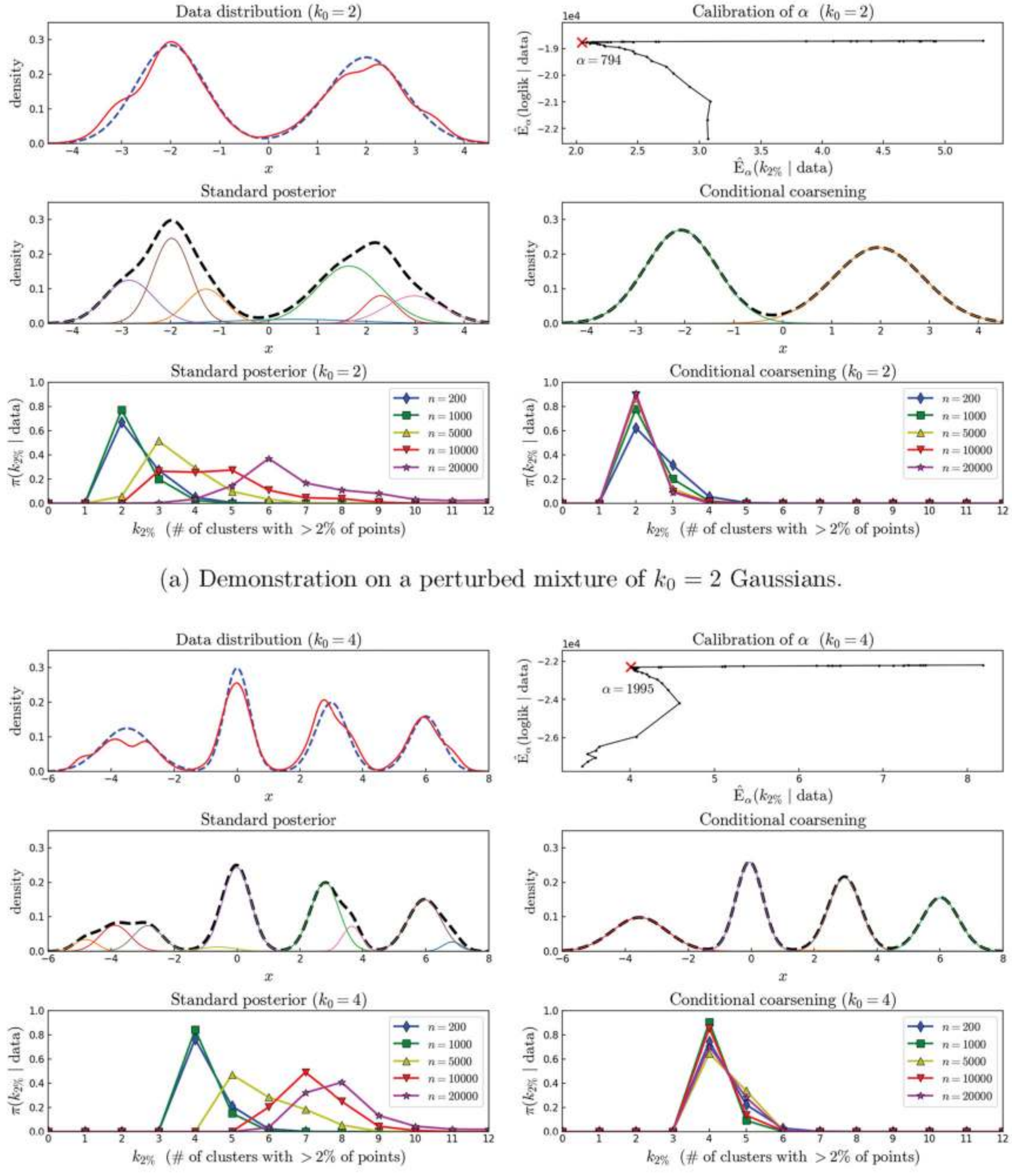


Figure 2.

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



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

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

Figure 3.

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

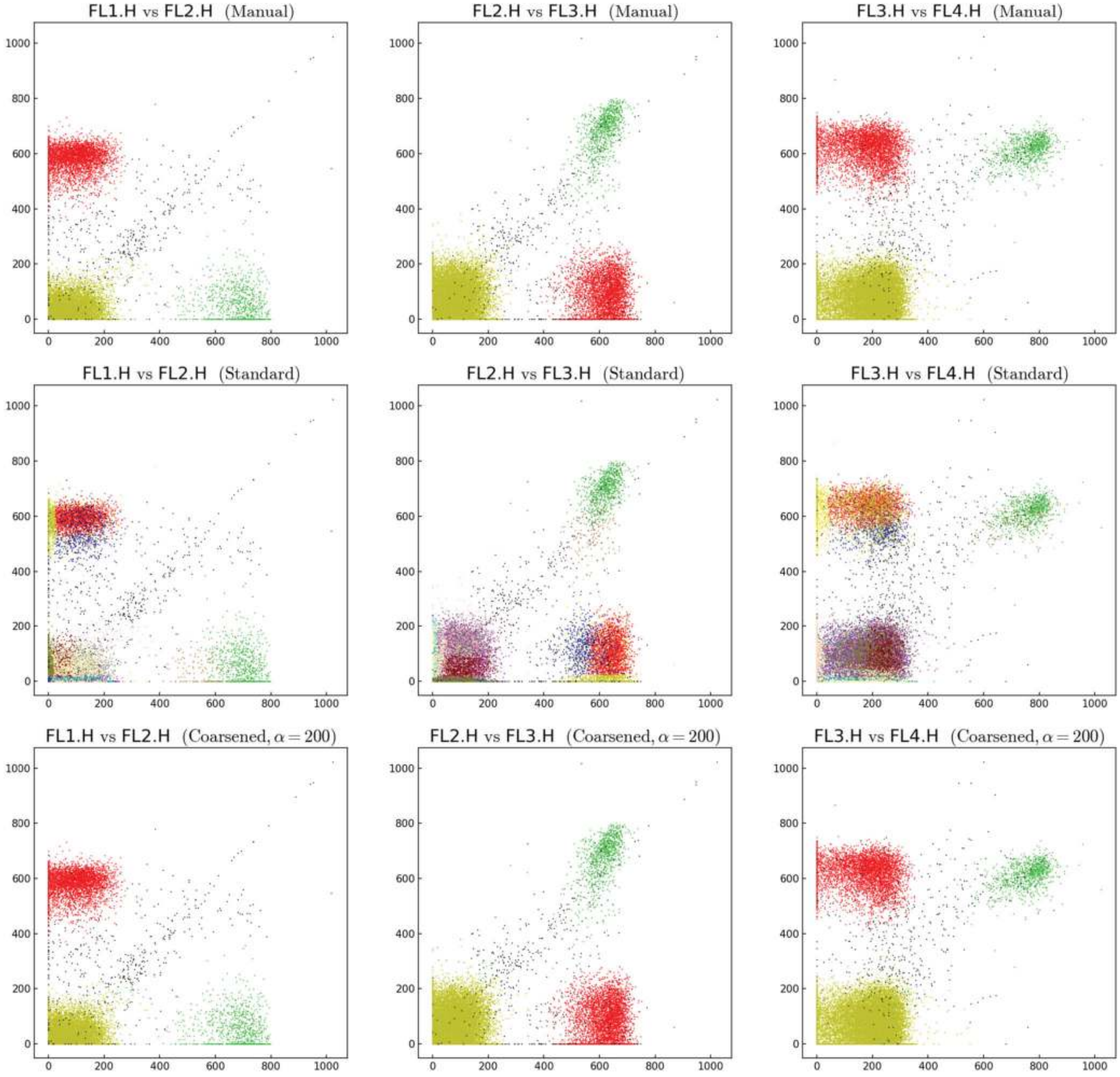


Figure 4.

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

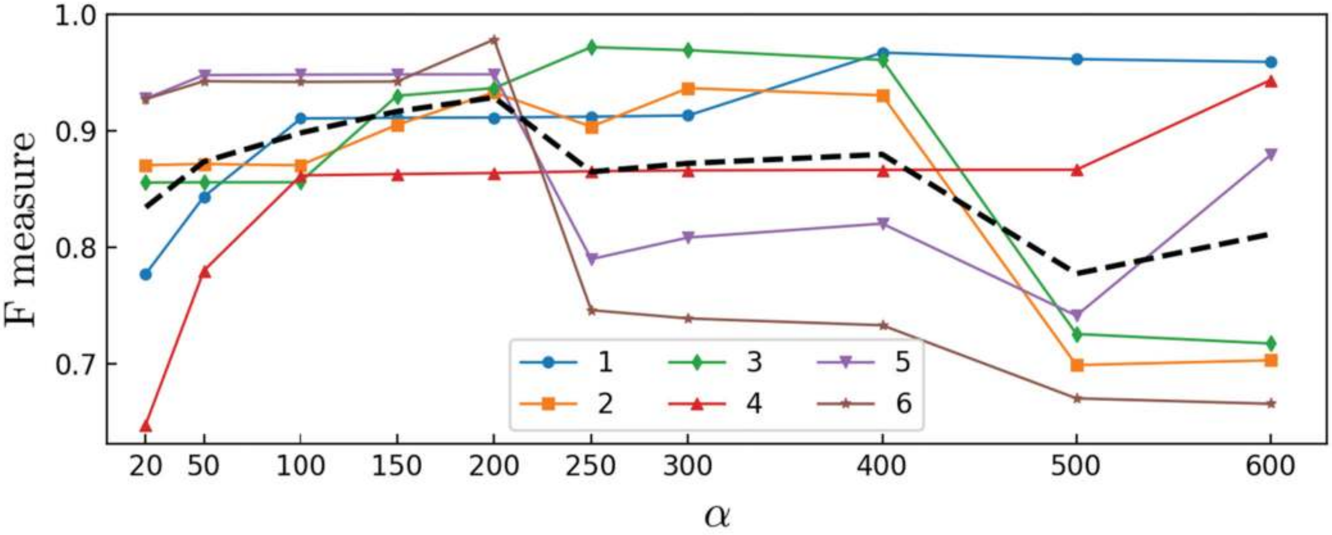


Figure 5.

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

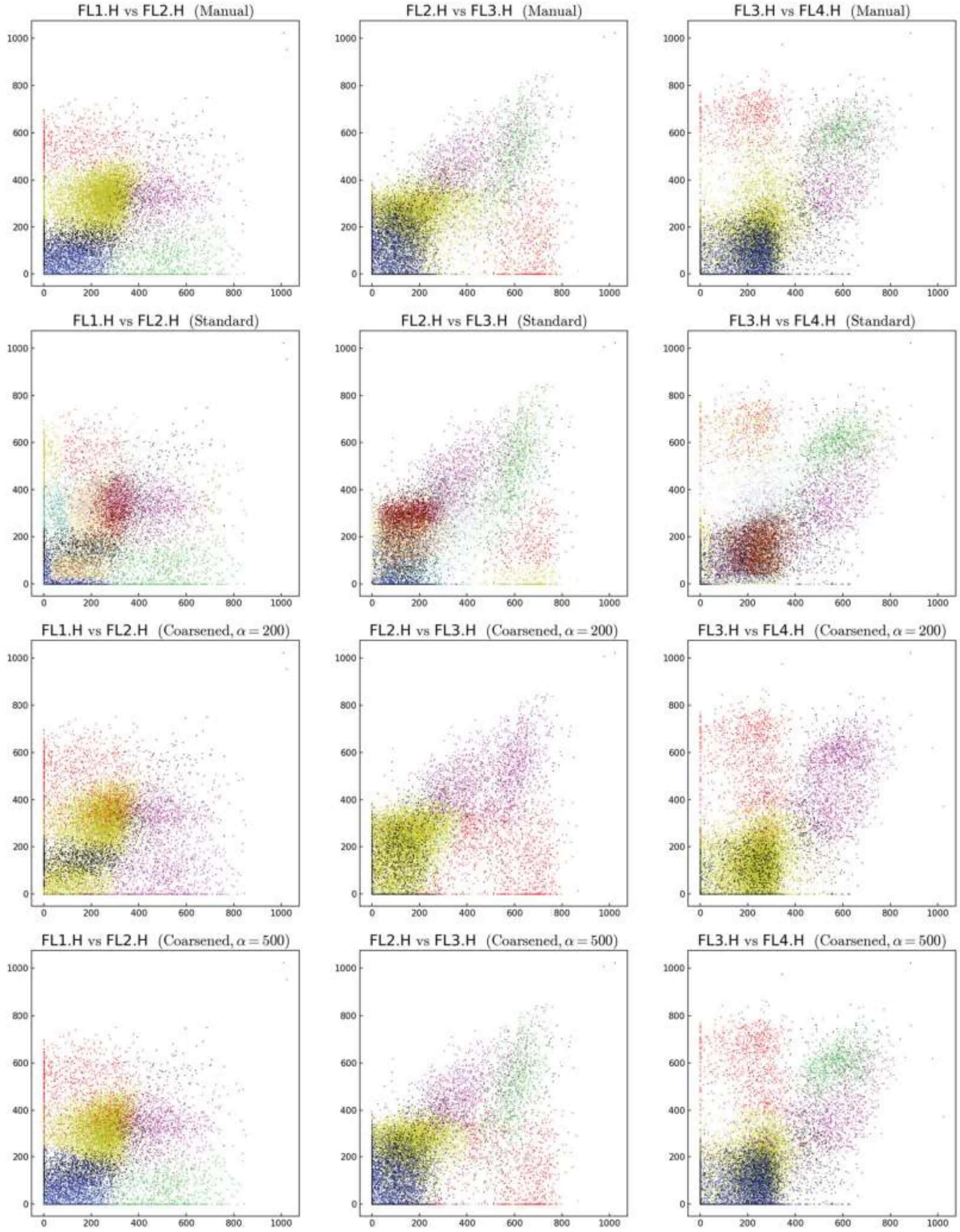


Figure 6.

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

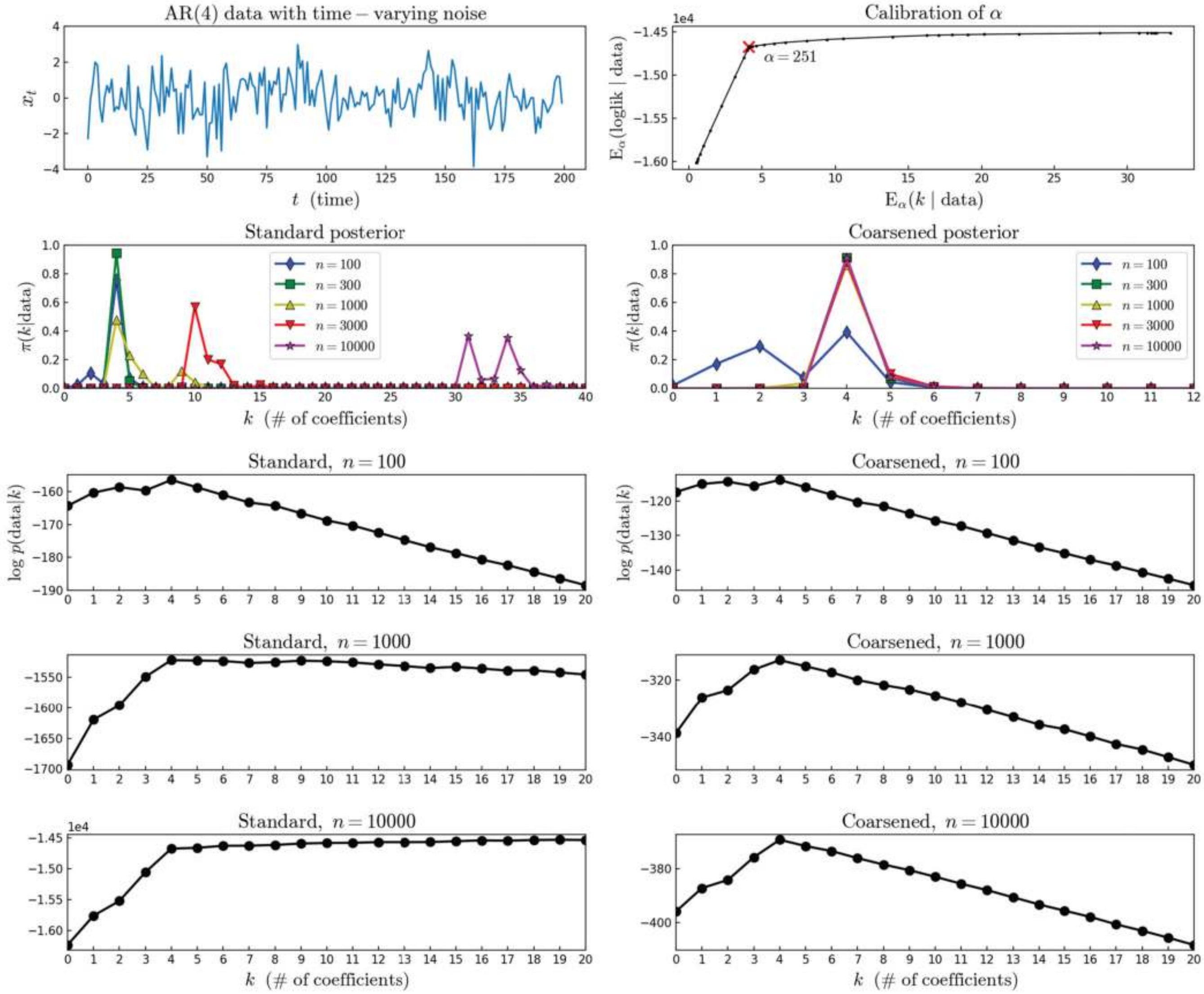


Figure 7.

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

Table 1

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

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