Robust Bayesian inference via coarsening

David Dunson

Joint work with Jeff Miller

Duke University
Department of Statistical Science

Wed April 9, 2025

Outline

1. Motivation
2. Our proposal: Coarsened posterior
3. Examples
   - Toy example: Bernoulli trials
   - Mixture models with an unknown number of components
4. Theory
5. More examples
   - Autoregressive models of unknown order
   - Variable selection in linear regression

Outline

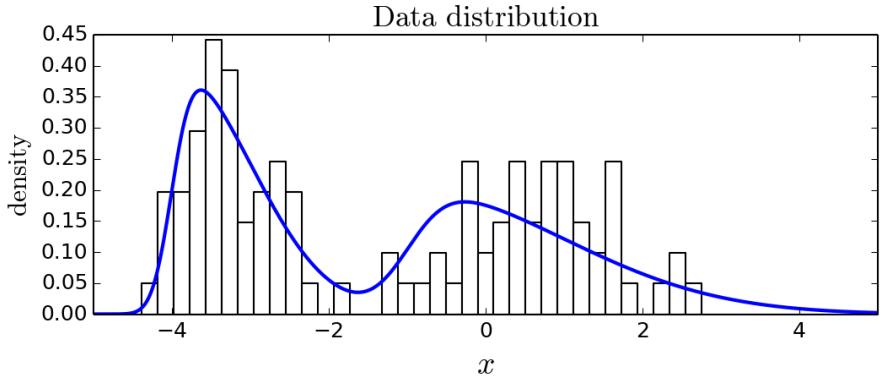
- In standard Bayesian inference, it is assumed that the model is correct.
- However, small violations of this assumption can have a large impact, and unfortunately, "all models are wrong."

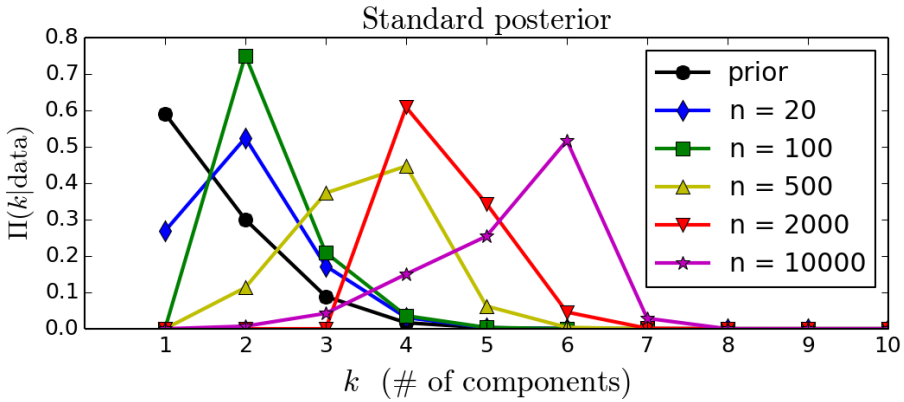
- In standard Bayesian inference, it is assumed that the model is correct.
- However, small violations of this assumption can have a large impact, and unfortunately, "all models are wrong."

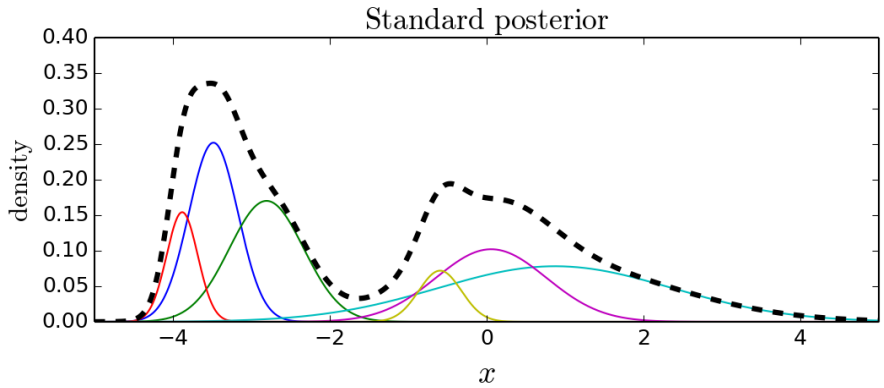
- In standard Bayesian inference, it is assumed that the model is correct.
- However, small violations of this assumption can have a large impact, and unfortunately, "all models are wrong."

- Is it possible to draw coherent inferences from a misspecified model?
- Can this be done in a computationally-tractable way?
- In the context of model averaging and Bayesian nonparametrics, can we be tolerant of models that are "close enough"?

Example: Mixture models



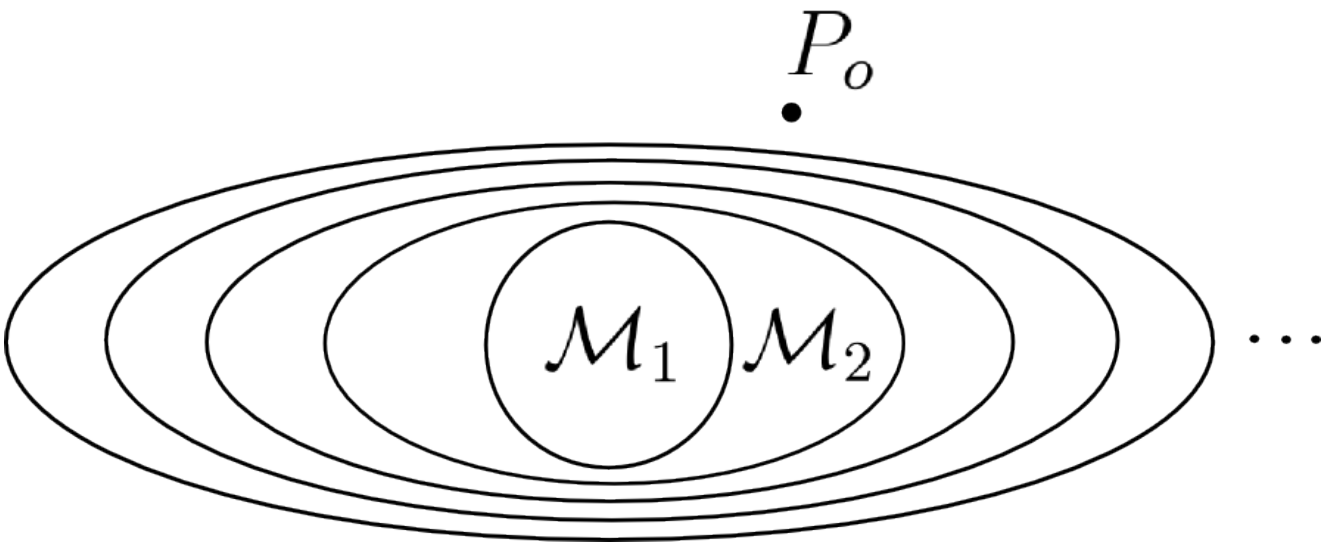




- Mixtures are often used for clustering.
- But if the data distribution is not exactly a mixture from the assumed family, the posterior will tend to introduce more and more clusters as n grows, in order to fit the data.
- As a result, the interpretability of the clusters may break down.

Somewhat more generally

Suppose we have a nested sequence of models $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \cdots$, but the distribution of the observed data, P_o, doesn't belong to any $\mathcal{M}_k$.



We seek an approach that tolerates models that are "close enough" to P_o.

Wait, if the model is wrong, why not just fix it?

- This is often impractical for a number of reasons.
  - insufficient insight into the data generating process
  - time and effort to design model + algorithms, and develop theory
  - slower and more complicated to do inference
  - complex models are less likely to be used in practice

Wait, if the model is wrong, why not just fix it?

- This is often impractical for a number of reasons.
  - insufficient insight into the data generating process
  - time and effort to design model + algorithms, and develop theory
  - slower and more complicated to do inference
  - complex models are less likely to be used in practice
- Further, a simple model may be more appropriate, even if wrong.
  - If there is a lack of fit, it may be due to contamination.
  - Many models are idealizations that are known to be inexact, but have interpretable parameters that provide insight into the questions of interest.
  - Often, the purpose of a model is to provide a lens through which to understand the data, rather than just fitting it.

There are many reasons to prefer simple, interpretable, efficient models. But we need a way to do inference that is robust to misspecification.

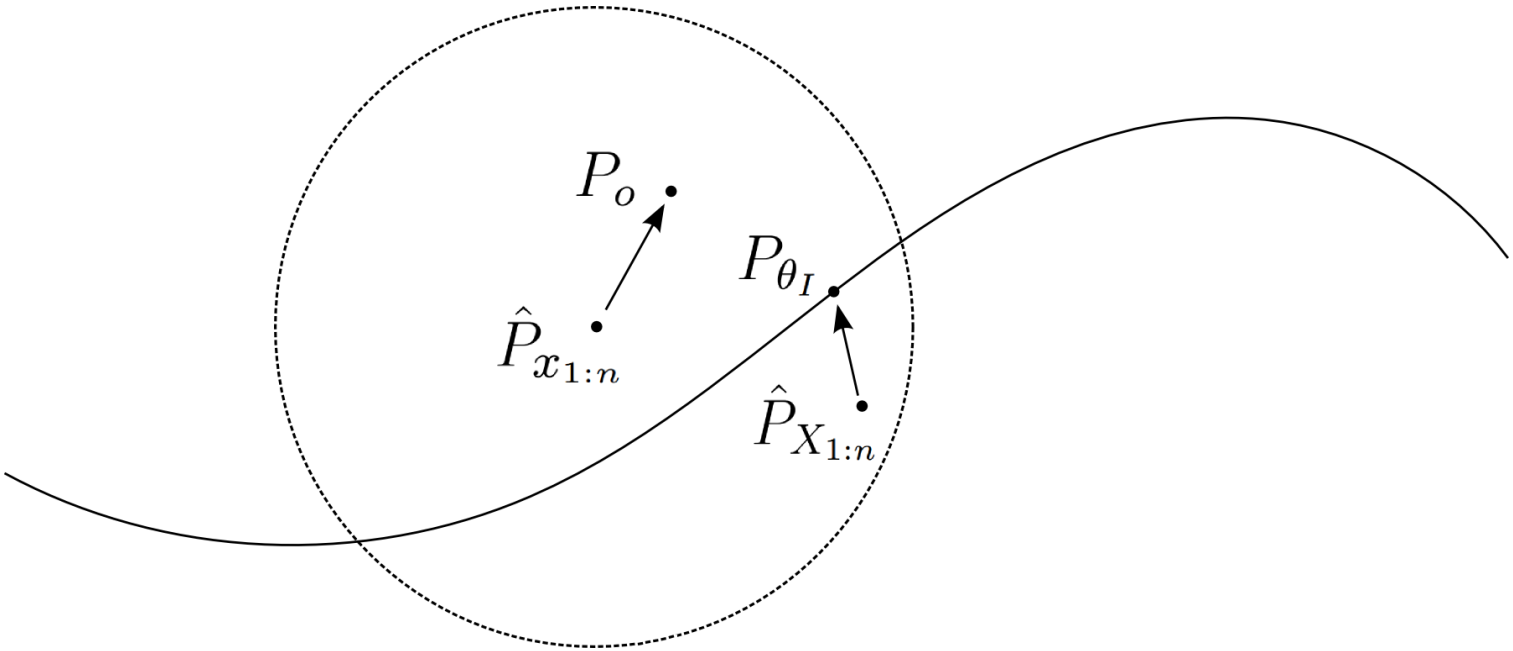
Some work on Bayesian robustness

- Gibbs posteriors (Jiang and Tanner, 2008)
- nonparametric approaches (Rodríguez and Walker, 2014)
- disparity-based posteriors (Hooker and Vidyashankar, 2014)
- learning rate adjustment (Grünwald and van Ommen, 2014)
- restricted posteriors (Lewis, MacEachern, and Lee, 2014)
- neighborhood methods (Liu and Lindsay, 2009)

There are interesting connections between these methods and ours, but our approach seems to be novel.

Outline

Our proposal: Coarsened posterior



- Assume a model $\{P_\theta : \theta \in \Theta\}$ and a prior $\pi(\theta)$.
- Suppose $\theta_I \in \Theta$ represents 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.
- Suppose $X_1, \ldots, X_n$ i.i.d. $\sim P_{\theta_I}$ are unobserved *idealized data*.
- However, the *observed data* $x_1, \ldots, x_n$ 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 $d(\cdot, \cdot)$.

Our proposal: Coarsened posterior

- If there were no corruption, then we should use the standard posterior

$$\pi(\theta \mid X_{1:n} = x_{1:n}).$$

- However, due to the corruption this would clearly be incorrect.
- Instead, a natural Bayesian approach would be to condition on what is known, giving us the *coarsened posterior* or *c-posterior*,

$$\pi(\theta \mid d(\hat{P}_{X_{1:n}}, \hat{P}_{x_{1:n}}) < R).$$

- Since R may be difficult to choose *a priori*, put a prior on it: $R \sim H$.
- More generally, consider

$$\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$$

where $d_n(X_{1:n}, x_{1:n}) \geq 0$ is some measure of the discrepancy between $X_{1:n}$ and $x_{1:n}$.

Connection with ABC

- The c-posterior $\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$ is mathematically equivalent to the approximate posterior resulting from *approximate Bayesian computation* (ABC).
- Tavaré et al. (1997), Marjoram et al. (2003), Beaumont et al. (2002), Wilkinson (2013)
- However, there are some crucial distinctions:
  - ABC is for intractable likelihoods, not robustness.
  - We assume the likelihood is tractable, facilitating computation.
  - For us, the c-posterior is an asset, not a liability.

Pros/cons of c-posteriors

Pros

- Robustness to small departures from the model.
  - Inherits the continuity properties of the chosen statistical distance.
- Coherent Bayesian inference based on limited information.
  - Use the same model, but conditioned on a different event than usual.
- Efficient computation in the case of relative entropy.
  - C-posterior can be approximated by simply tempering the likelihood.
- Simple asymptotic form, facilitating computation and analysis.

Cons

- Sometimes less concentrated than one would like.
  - e.g., if there is less misspecification than expected.

Relative entropy c-posteriors

- There are many possible choices of statistical distance . . .
  - e.g., Kolmogorov–Smirnov, Wasserstein, maximum mean discrepancy, various divergences
- . . . but relative entropy works out exceptionally nicely.
- Suppose $d_n(X_{1:n}, x_{1:n})$ is a consistent estimator of $D(p_o \| p_\theta)$ when $X_i \overset{\text{iid}}{\sim} p_\theta$ and $x_i \overset{\text{iid}}{\sim} p_o$.
- When $R \sim \text{Exp}(\alpha)$, we have the *power posterior* approximation,

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

  where $\zeta_n = (1/n)/(1/n + 1/\alpha)$.
- The power posterior enables inference using standard techniques:
  - analytical solutions in the case of conjugate priors
  - Gibbs sampling when using conditionally-conjugate priors
  - Metropolis–Hastings MCMC, more generally

How to choose the "precision" α?

- Strategy #1. Set the mean neighborhood size $\mathbb{E}R = 1/\alpha$ to match the amount of misspecification we expect.
- Strategy #2. Rule of thumb: to be robust to perturbations that would require at least N samples to distinguish, set $\alpha \approx N$.
- Strategy #3. Consider a range of α values, for sensitivity analysis or exploratory analysis.

Some work on power likelihoods

- *Power likelihoods* of the form $\prod_{i=1}^{n} p_\theta(x_i)^\zeta$ have been used previously.
- Usually, this is done for reasons completely unrelated to robustness.
  - marginal likelihood approximation (Friel and Pettitt, 2008)
  - improved MCMC mixing (Geyer, 1991)
  - consistency in nonparametrics (Walker and Hjort, 2001; Zhang, 2006a)
  - discounting historical data (Ibrahim and Chen, 2000)
  - objective Bayesian model selection (O'Hagan, 1995)
- Grünwald and van Ommen (2014) found that a power posterior improves robustness.
- However, the form of power we use, and its theoretical justification, seem novel.

Outline

Toy example: Bernoulli trials

- Model: $X_1, \ldots, X_n | \theta$ i.i.d. $\sim \text{Bernoulli}(\theta)$
- Interested in testing $\text{H}_0 : \theta = 1/2$ versus $\text{H}_1 : \theta \neq 1/2$.
- Prior: $\pi(\text{H}_0) = \pi(\text{H}_1) = 1/2$, and $\theta | \text{H}_1 \sim \text{Uniform}(0, 1)$.
- Standard posterior:

$$\pi\big(\text{H}_0 \,\big|\, X_{1:n} = x_{1:n}\big) = 1/\big(1 + 2^n B(1 + n\overline{x},\, 1 + n(1 - \overline{x}))\big)$$

- Suppose, however, the observed data $x_1, \ldots, x_n$ is slightly corrupted.
- Coarsened posterior:

$$\pi\big(\text{H}_0 \,\big|\, D(\hat{p}_x || \hat{p}_X) < R\big) \approx 1/\big(1 + 2^{\alpha_n} B(1 + \alpha_n \overline{x},\, 1 + \alpha_n(1 - \overline{x}))\big)$$

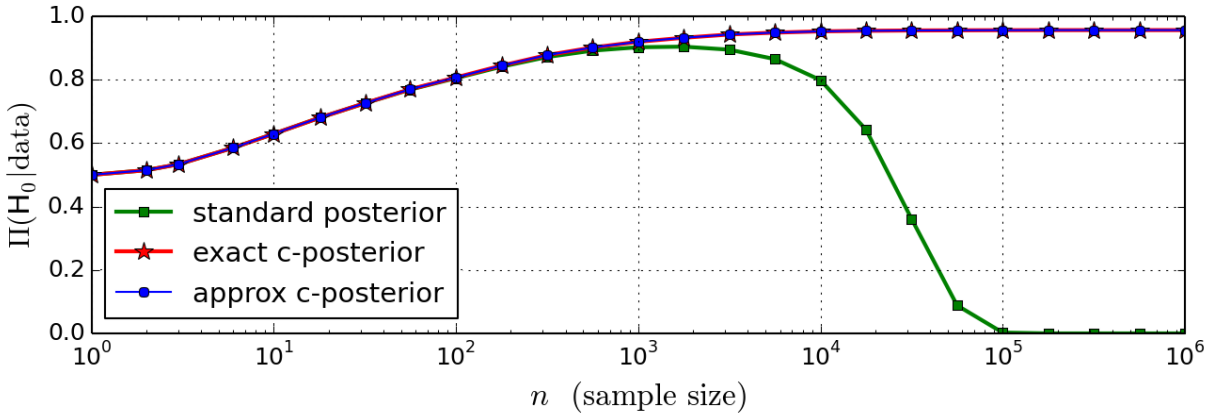
  where $\alpha_n = 1/(1/n + 1/\alpha)$ and $R \sim \text{Exp}(\alpha)$.
- What to choose for α?

Toy example: Bernoulli trials

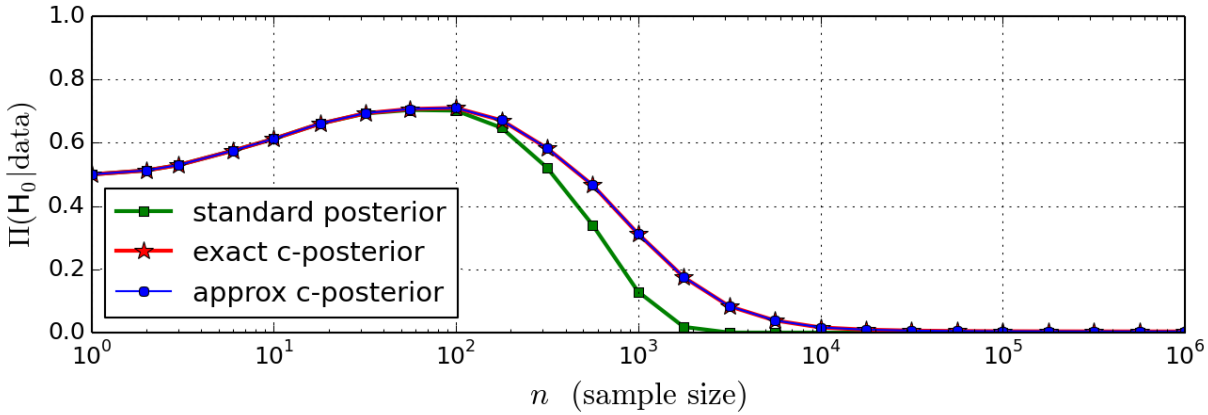
- What to choose for α?
- Use strategy #1: Set the mean neighborhood size $\mathbb{E}R = 1/\alpha$ to match the amount of misspecification we expect.
- For example, suppose we expect the misspecification to affect $\bar{x}$ by no more than, say, $\varepsilon = 0.02$ when $\theta = 1/2$.
- By the chi-squared approximation to relative entropy, we have $D(\hat{p}_x || \hat{p}_X) \approx 2|\bar{x} - \bar{X}|^2$ when $\bar{x}$ and $\bar{X}$ are near $1/2$.
- This suggests choosing $\alpha = 1/(2\varepsilon^2) = 1/(2 \cdot 0.02^2) = 1250$.

Toy example: Bernoulli trials

Suppose H_0 is true, but $x_1, \ldots, x_n$ are corrupted and behave like $\mathrm{Bernoulli}(0.51)$ samples. The c-posterior is robust to this, but the standard posterior is not.



What if the departure from H_0 is significantly larger than our chosen tolerance of $\varepsilon = 0.02$, e.g., if $x_1, \ldots, x_n$ are $\mathrm{Bernoulli}(0.56)$ samples? Does the c-posterior more strongly favor H_1 in such cases, as it should? Indeed, it does.



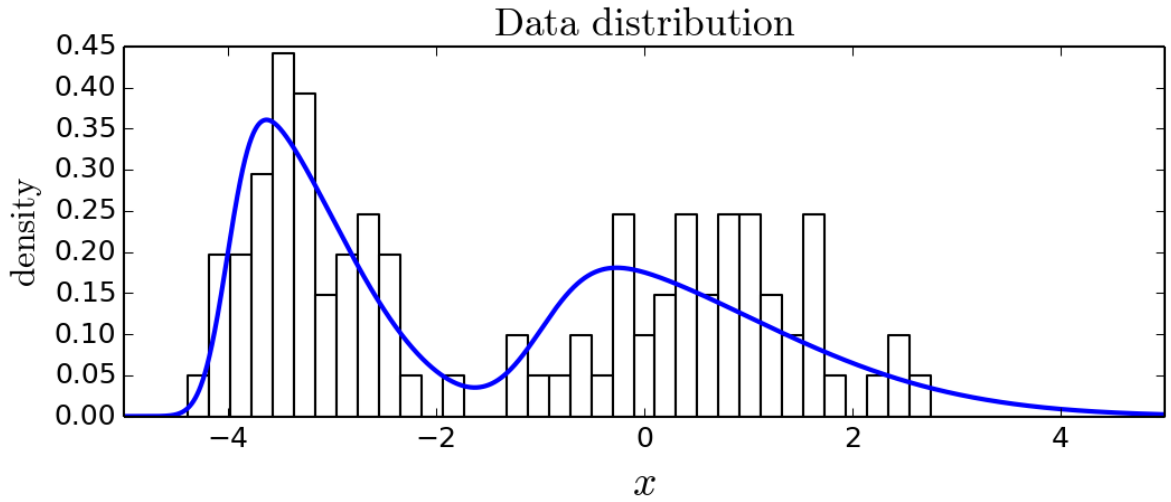
Example: Gaussian mixture with a prior on k

- Model: $X_1, \ldots, X_n | k, w, \varphi$ i.i.d. $\sim \sum_{i=1}^{k} w_i f_{\varphi_i}(x)$
- Prior $\pi(k, w, \varphi)$ on # of components k, weights w, and params φ.
- Relative entropy c-posterior is approximated by the power posterior,

$$\pi\big(k, w, \varphi \,\big|\, d_n(X_{1:n}, x_{1:n}) < R\big) \propto\!\!\!\propto \pi(k, w, \varphi) \prod_{j=1}^{n} \Big(\sum_{i=1}^{k} w_i f_{\varphi_i}(x_j) \Big)^{\zeta_n}$$

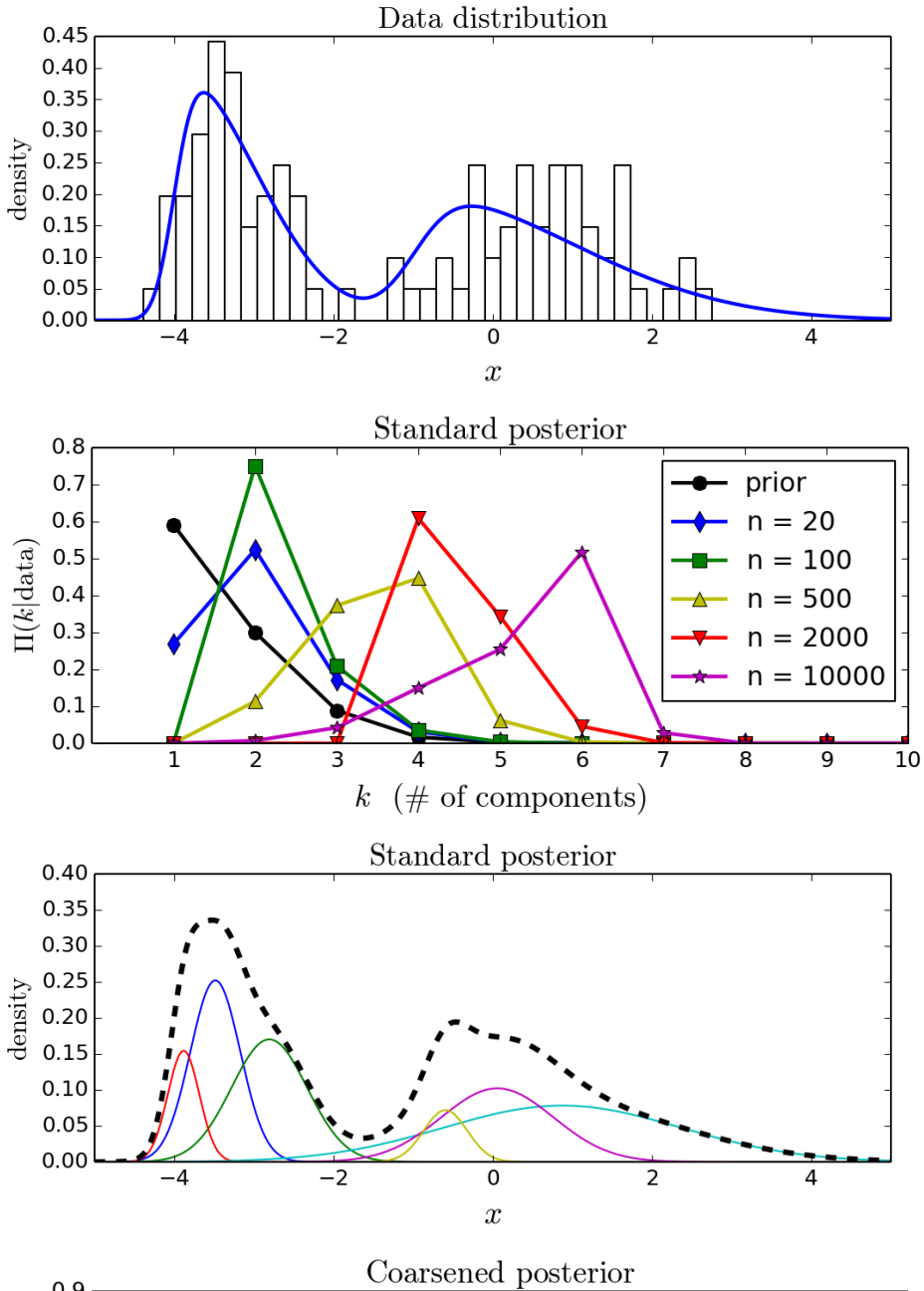
  where $\zeta_n = (1/n)/(1/n + 1/\alpha)$.
- Could use Antoniano-Villalobos and Walker (2013) algorithm or RJMCMC (Green, 1995). For simplicity, we reparametrize in a way that allows the use of plain-vanilla Metropolis–Hastings.

Gaussian mixture applied to skew-normal mixture data

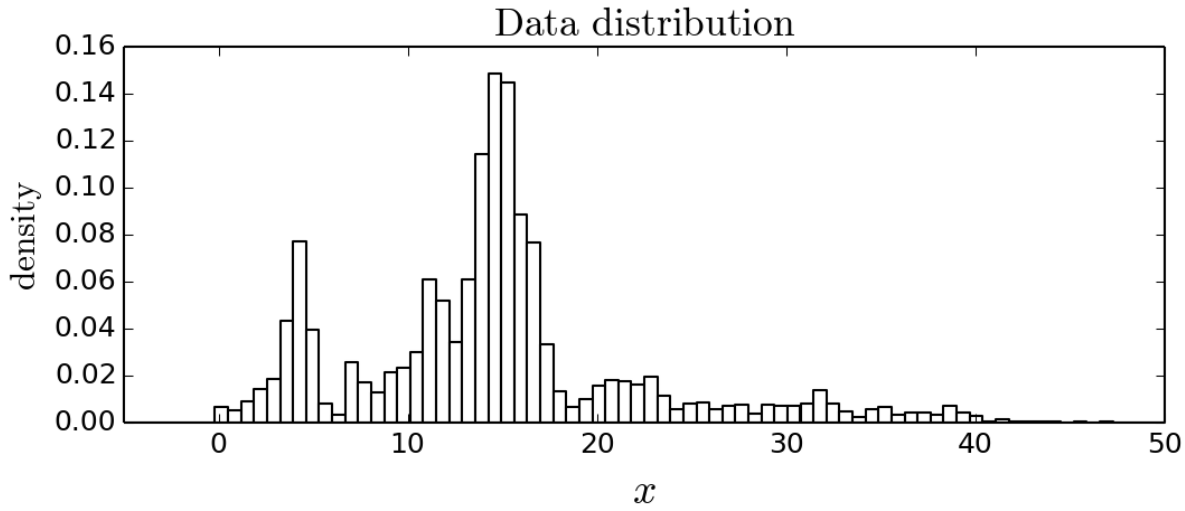


- Data: $x_1, \ldots, x_n$ i.i.d. $\sim \frac{1}{2}\mathcal{SN}(-4, 1, 5) + \frac{1}{2}\mathcal{SN}(-1, 2, 5)$, where $\mathcal{SN}(\xi, s, a)$ is the skew-normal distribution with location ξ, scale s, and shape a (Azzalini and Capitanio, 1999).
- Use strategy #2: Choose $\alpha = 100$, to be robust to perturbations to P_o that would require at least 100 samples to distinguish, roughly speaking.

Gaussian mixture applied to skew-normal mixture data

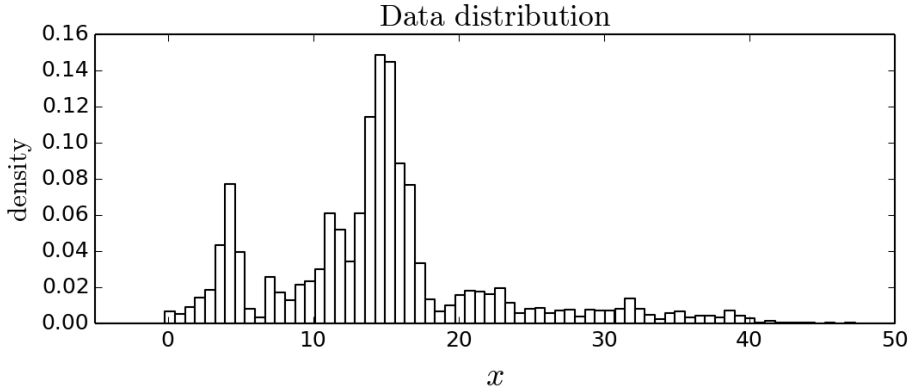


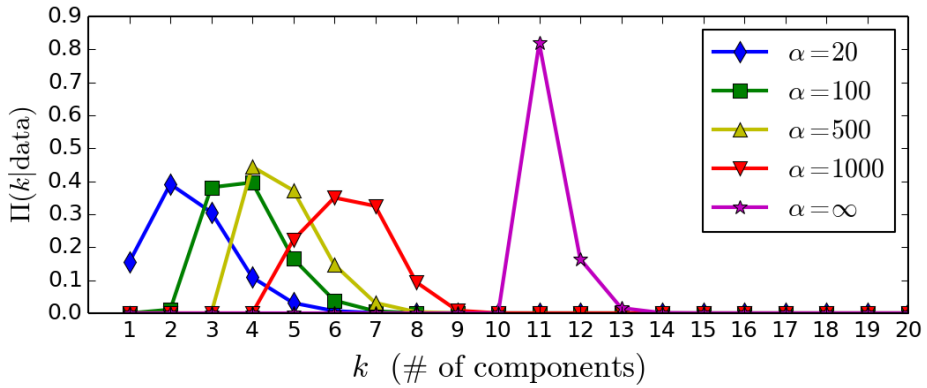
Velocities of galaxies in the Shapley supercluster

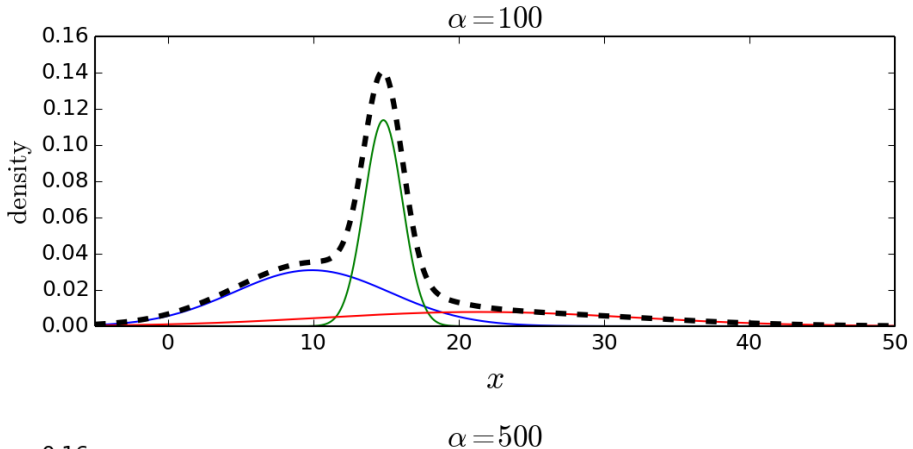


- Velocities of 4215 galaxies in a large concentration of gravitationally-interacting galaxies (Drinkwater et al., 2004).
- Gaussian mixture assumption is probably wrong.
- Use strategy #3: By considering a range of α values, we can explore the data at varying levels of precision.

Velocities of galaxies in the Shapley supercluster







Outline

Theory

We establish two main theoretical results:

1. the asymptotic form of c-posteriors as $n \to \infty$, and
2. robustness of c-posteriors to perturbations of the data distribution.

Consider the model

$$\begin{aligned}
&\boldsymbol{\theta} \sim \Pi \\
&X_1, \ldots, X_n | \boldsymbol{\theta} \text{ i.i.d.} \sim P_{\boldsymbol{\theta}} \\
&R \in [0, \infty) \text{ independently of } \boldsymbol{\theta}, X_{1:n}.
\end{aligned}$$

Suppose the observed data $x_1, \ldots, x_n$ are sampled i.i.d. from some P_o.

Theory: Asymptotic form

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

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

Theorem (Asymptotic form of c-posteriors)

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

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

and in fact,

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

for any $h \in L^1(\Pi)$.

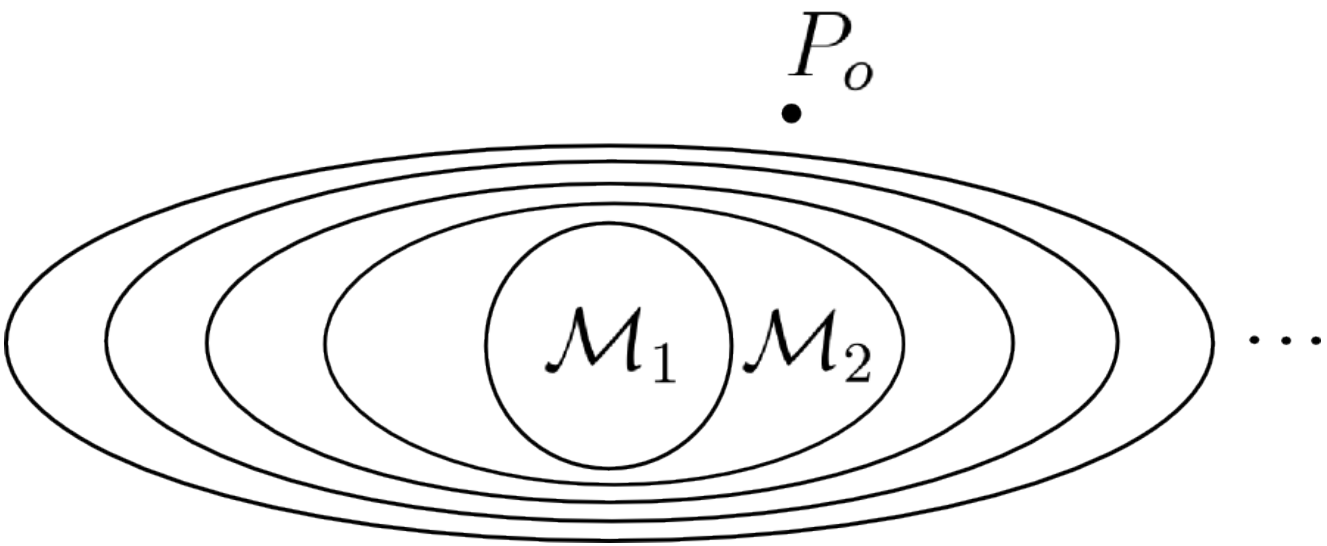
Theory: Lack of robustness of the standard posterior

- The standard posterior can be strongly affected by small changes to the observed data distribution P_o, particularly when doing model inference.
- Roughly,

$$\begin{aligned}\pi(\theta \mid x_{1:n}) &\propto \exp\Big(\sum_{i=1}^{n} \log p_\theta(x_i)\Big)\pi(\theta) \\ &\approx \exp\big(n \textstyle\int p_o \log p_\theta\big)\pi(\theta) \\ &\propto \exp(-nD(p_o \| p_\theta))\pi(\theta).\end{aligned}$$

- Due to the n in the exponent, even a slight change to P_o can dramatically change the posterior.

Theory: Lack of robustness of the standard posterior



Theory: Robustness

- Roughly, robustness means that small changes to the data distribution result in small changes to the resulting inferences.
- This can be formalized in terms of continuity with respect to P_o.
- The asymptotic c-posterior inherits the continuity properties of whatever distance $d(\cdot, \cdot)$ is used to define it.

Theorem (Robustness of c-posteriors)

If $P_1, P_2, \ldots$ *such that* $d(P_\theta, P_m) \xrightarrow[m \to \infty]{} d(P_\theta, P_o)$ *for* Π*-almost all* $\theta \in \Theta$, *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,*

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

Outline

Example: Autoregressive $\mathrm{AR}(k)$ model with a prior on k

- Model: $X_t = \sum_{\ell=1}^{k} \theta_\ell X_{t-\ell} + \varepsilon_t$ where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- Prior $\pi(k)$ on k, and $\theta_1, \ldots, \theta_k | k \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_0^2)$. Assume σ^2 known.
- For time series, a natural choice of distance is relative entropy rate.
- The c-posterior based on relative entropy rate estimates $d_n(X_{1:n}, x_{1:n})$ is again approximated by a power posterior,

$$\propto p(x_{1:n}|\theta, k)^{\zeta_n} \pi(\theta|k)\pi(k).$$

- This leads to the coarsened marginal likelihood for k,

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

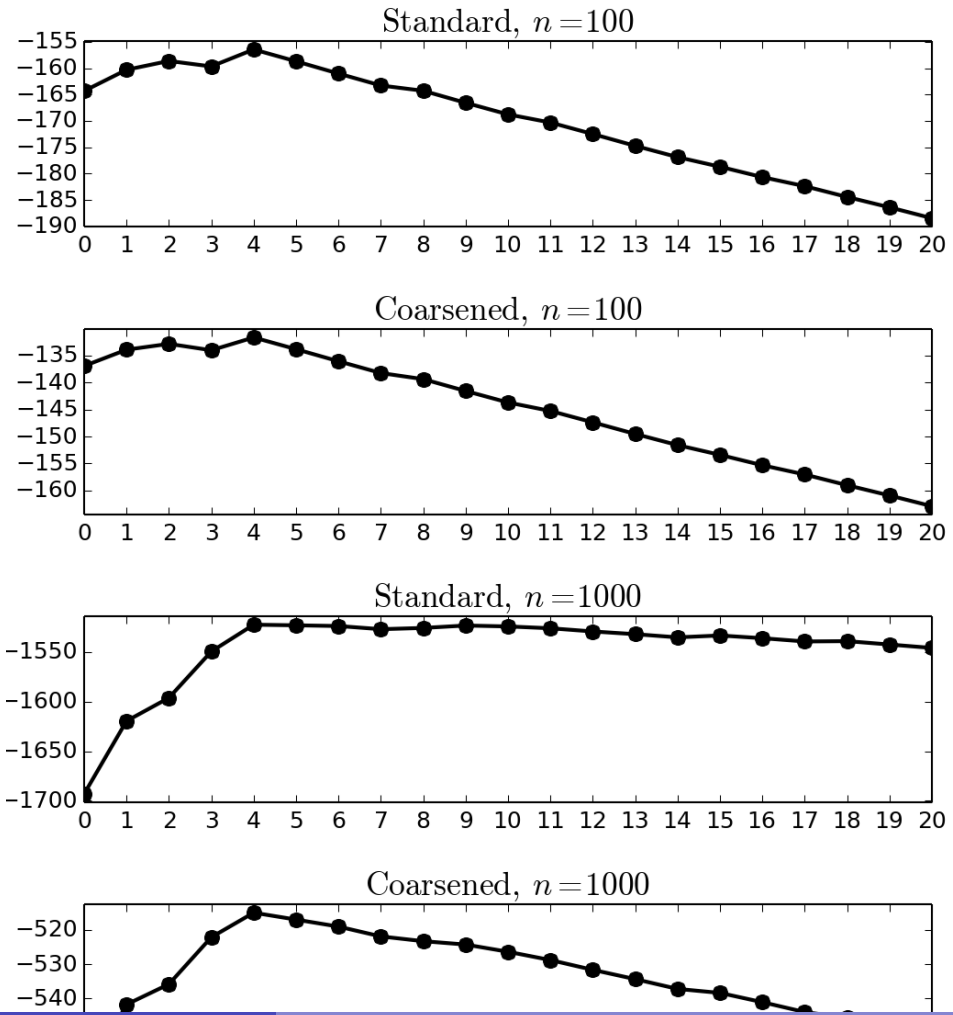
  where $\zeta_n = (1/n)/(1/n + 1/\alpha)$.
- This can be computed analytically, since $\theta|k$ has been given a conjugate prior.

Suppose the data is close to $\mathrm{AR}(4)$ but has time-varying noise:

$$x_t = \tfrac{1}{4}(x_{t-1} + x_{t-2} - x_{t-3} + x_{t-4}) + \varepsilon_t + \tfrac{1}{2}\sin t$$

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

Log of marginal likelihood



Example: Variable selection in linear regression

- Spike-and-slab model:

$$W \sim \text{Beta}(1, 2p)$$
$$\beta_j \sim \mathcal{N}(0, \sigma_0^2) \text{ with probability } W, \text{ otherwise } \beta_j = 0, \text{ for } j = 1, \ldots, p$$
$$\sigma^2 \sim \text{InvGamma}(a, b)$$
$$Y_i | \beta, \sigma^2 \sim \mathcal{N}(\beta^{\mathsf{T}} x_i, \sigma^2) \text{ independently for } i = 1, \ldots, n.$$

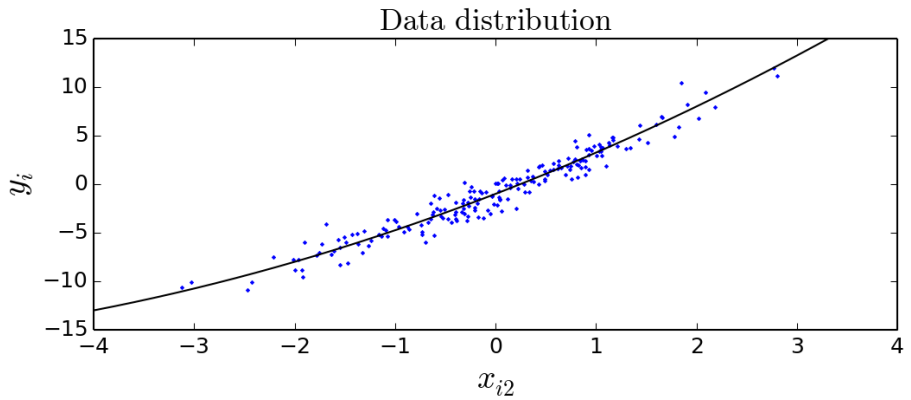
- For regression, a natural choice of statistical distance is conditional relative entropy. Again, this leads to a power posterior approximation to the c-posterior:

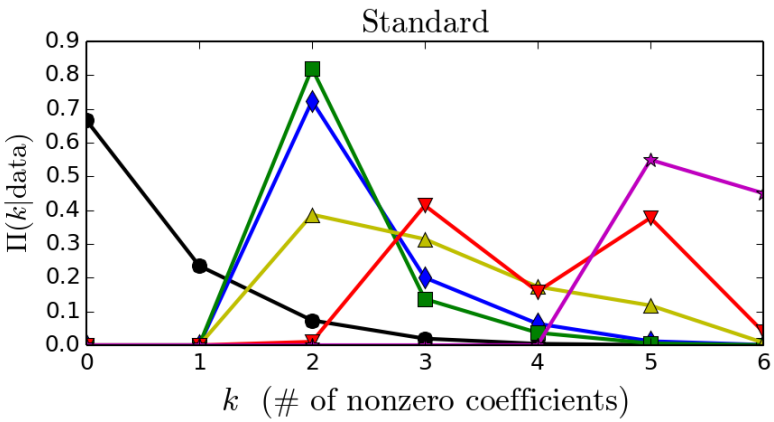
$$\pi\left(\beta, \sigma^2 \,\middle|\, d_n(Y_{1:n}, y_{1:n}) < R\right) \propto\!\!\!\propto \pi(\beta, \sigma^2) \prod_{i=1}^{n} p(y_i | x_i, \beta, \sigma^2)^{\zeta_n}.$$

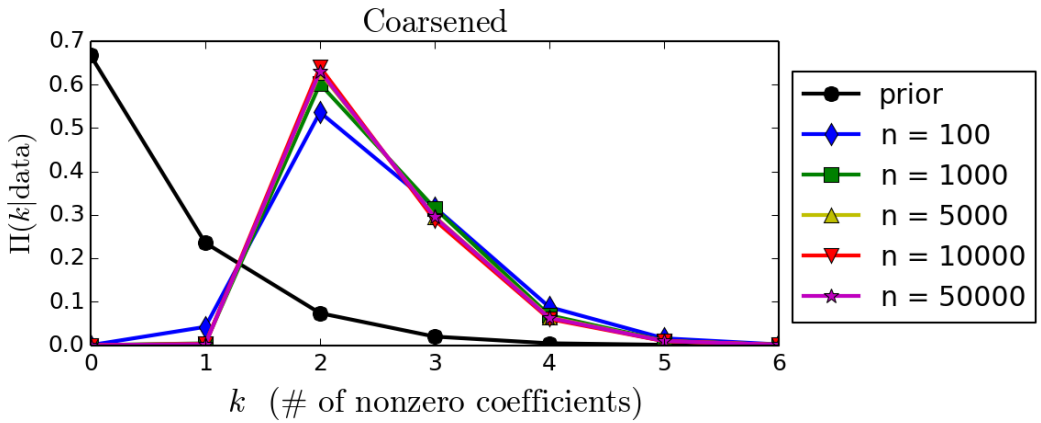
- Since we are using conditionally-conjugate priors, the full conditionals can be derived in closed-form, and we can use Gibbs sampling.

Simulation example for variable selection

- Covariates: $x_{i1} = 1$ to accomodate constant offset, and $x_{i2}, \ldots, x_{i6}$ distributed according to a multivariate skew-normal distribution.
- $y_i = -1 + 4(x_{i2} + \frac{1}{16}x_{i2}^2) + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.
- Set $\alpha = 50$, using knowledge of the true amount of misspecification.

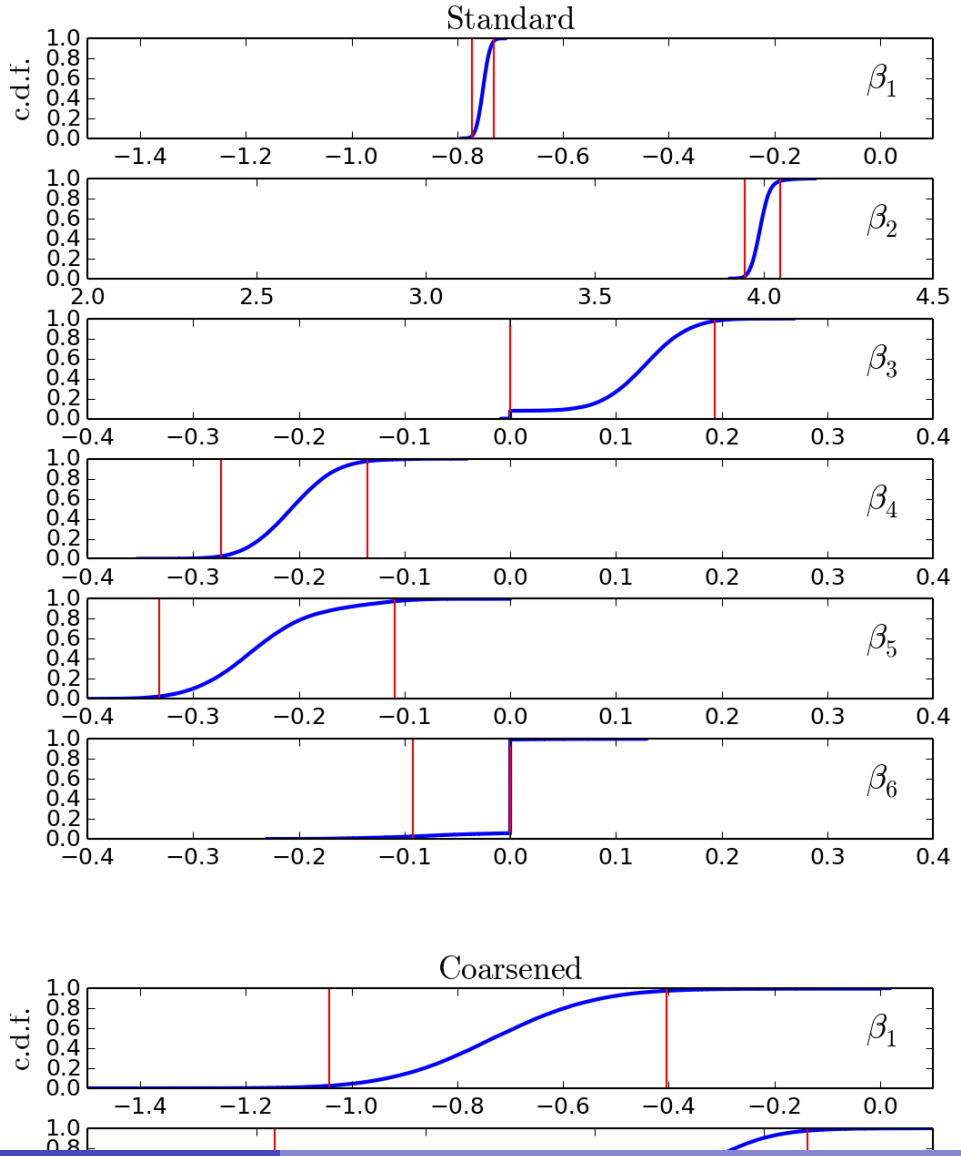






Simulation example for variable selection

Posterior c.d.f. for each coefficient (blue), and 95% credible interval (red)



Modeling birthweight of infants

- Pregnancy data from the Collaborative Perinatal Project.
- We use a subset with $n = 2379$ subjects, and $p = 72$ covariates that are potentially predictive of birthweight.
  - e.g., body length, mother's weight, gestation time, cigarettes/day smoked by mother, previous pregnancy, etc.
- Not sure how much misspecification there is, so we explore a range of "precision" values α:

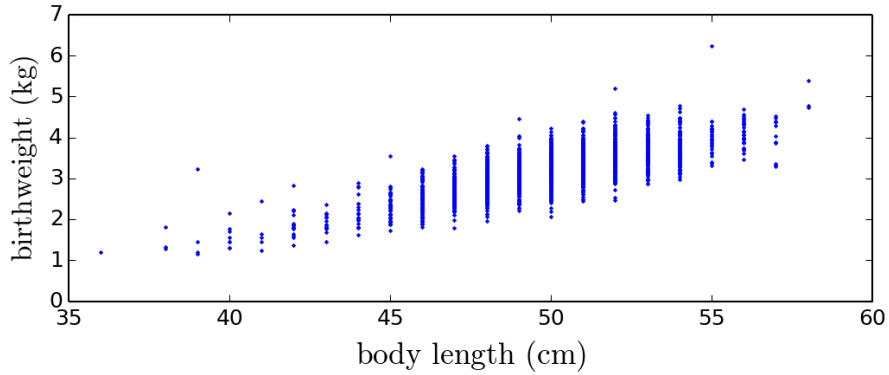
$$\alpha \in \{100, 500, 1000, 2000, \infty\}$$

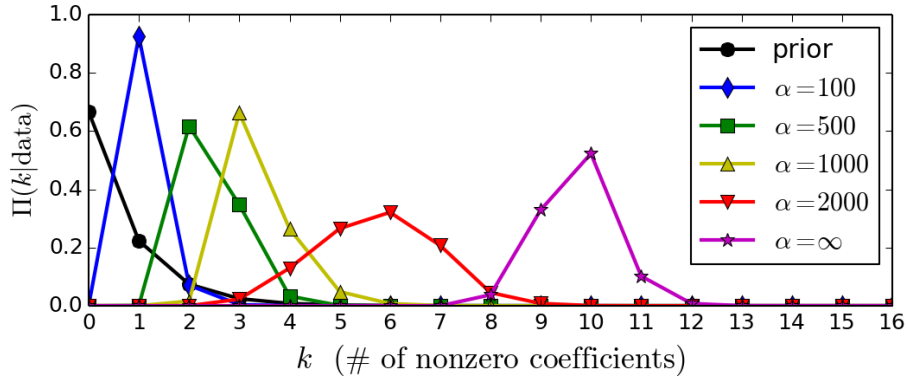
which corresponds roughly to contamination of magnitude

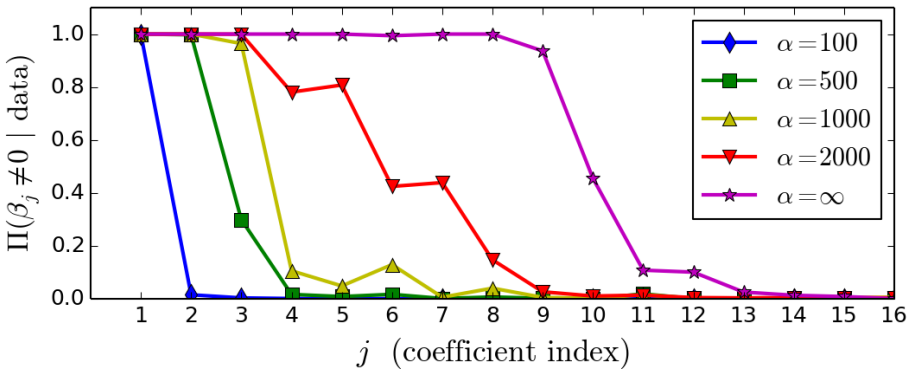
$$\delta \in \{0.045, 0.02, 0.015, 0.01, 0\} \text{ kilograms}$$

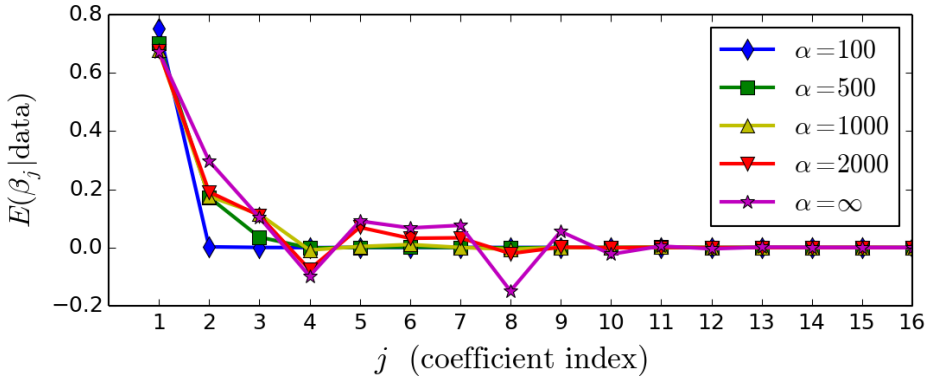
by the formula for the relative entropy between Gaussians.

Modeling birthweight of infants









Top variables: 1. Body length, 2. Mother's weight at delivery, 3. Gestation time, 4. African-American, etc.

Conclusion

The coarsened posterior (c-posterior) seems promising as a general approach to robust Bayesian inference.

Pros

- Robustness to small departures from the model.
  - Inherits the continuity properties of the chosen statistical distance.
- Coherent Bayesian inference based on limited information.
  - Use the same model, but conditioned on a different event than usual.
- Efficient computation in the case of relative entropy.
  - C-posterior can be approximated by simply tempering the likelihood.
- Simple asymptotic form, facilitating computation and analysis.

Cons

- Sometimes less concentrated than one would like.
  - e.g., if there is less misspecification than expected.