Robust Bayesian inference via coarsening

Jeff Miller

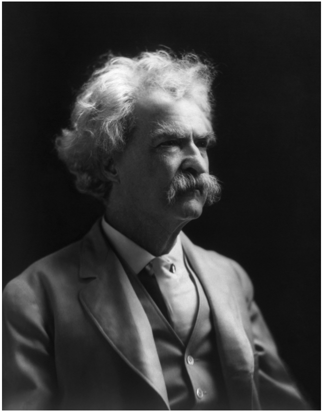
Joint work with David Dunson

Harvard University
Department of Biostatistics

Probability and Statistics Seminar, Boston University
March 15, 2018

"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so."

– attributed to Mark Twain



Outline

1. Motivation

2. Coarsened posterior, Power posterior

3. Examples
   - Toy example: Bernoulli trials
   - Mixture models and clustering
   - Autoregressive models of unknown order

4. Theory

Outline

1. Motivation
2. Coarsened posterior, Power posterior
3. Examples
    - Toy example: Bernoulli trials
    - Mixture models and clustering
    - Autoregressive models of unknown order
4. Theory

Motivation

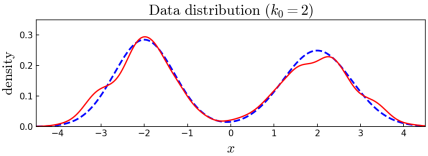
- In standard Bayesian inference, it is assumed that the model is correct.
- However, small violations of this assumption can have a large impact, and unfortunately, “all models are wrong.”

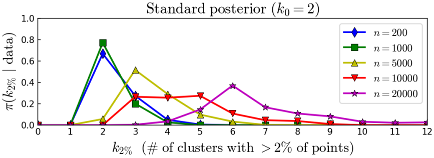
Motivation

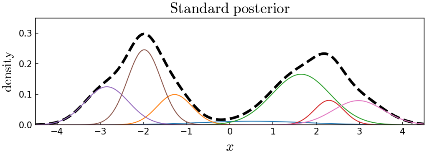
- In standard Bayesian inference, it is assumed that the model is correct.
- However, small violations of this assumption can have a large impact, and unfortunately, “all models are wrong.”

- 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: Perturbed mixture of Gaussians


Data distribution ($k_0 = 2$)

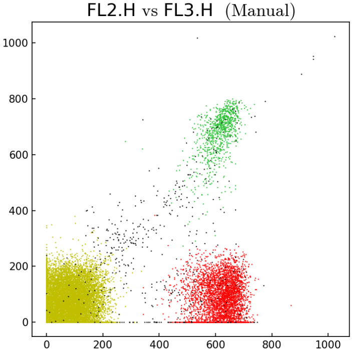

Standard posterior ($k_0 = 2$)

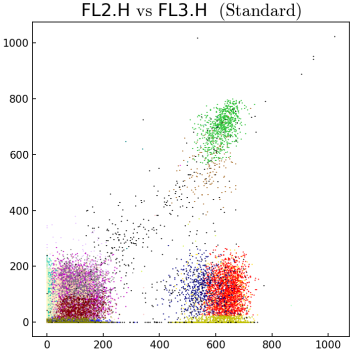

Standard posterior

- 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.

Example: Flow cytometry clustering

- Each sample has 3 to 20-dim measurements on 10K's of cells.
- Manual clustering is time-consuming and subjective.
- Multivariate Gaussian mixture yields too many clusters.
- Example: GvHD data from FLOWCAP-I.





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?

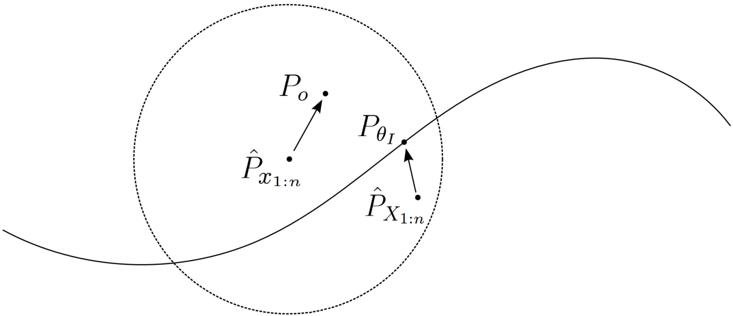
- This is often impractical for a number of reasons.
  - insufficient insight into the data generating process
  - time and effort to design model + algorithms, and develop theory
  - slower and more complicated to do inference
  - complex models are less likely to be used in practice
- Further, a simple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

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

Outline

1. Motivation
2. Coarsened posterior, Power posterior
3. Examples
    - Toy example: Bernoulli trials
    - Mixture models and clustering
    - Autoregressive models of unknown order
4. Theory

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 $x_{1:n}$.

Relative entropy c-posterior $\approx$ Power posterior

- 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 $R \sim \mathrm{Exp}(\alpha)$, 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 $\zeta_n = \alpha/(\alpha + n)$.
- 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

Previous 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)
- Recently, 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.

Recent 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.

How to choose the “precision” α?

- Ideally, use prior knowledge:
  - Set the mean neighborhood size $\mathbb{E}R = 1/\alpha$ to match the amount of misspecification we expect.
  - Or, to be robust to perturbations that would require at least N samples to distinguish, set $\alpha \approx N$.

- If no prior knowledge, can either:
  - Consider a range of α values, for sensitivity analysis or exploratory analysis.
  - Or, use our calibration curve technique — data-driven choice of α.

Outline

1. Motivation

2. Coarsened posterior, Power posterior

3. Examples
   - Toy example: Bernoulli trials
   - Mixture models and clustering
   - Autoregressive models of unknown order

4. Theory

Toy example: Bernoulli trials

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

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

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

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

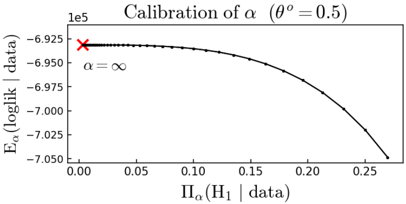
  where $\alpha_n = 1/(1/n + 1/\alpha)$ and $R \sim \mathrm{Exp}(\alpha)$.
- What to choose for α?

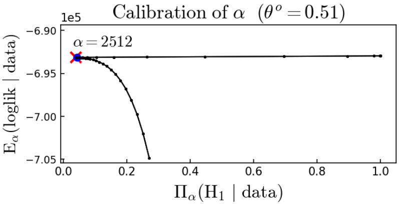
Choosing α? Prior knowledge approach

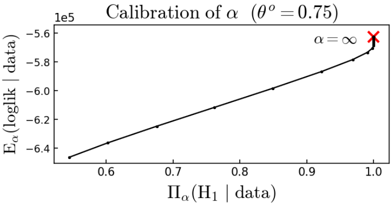
- Set the mean neighborhood size $\mathbb{E}R = 1/\alpha$ to match the amount of misspecification we expect.
- Suppose we expect the misspecification to affect $\bar{x}$ by no more than, say, $\varepsilon = 0.02$ when $\theta = 1/2$.
- By the chi-squared approximation to relative entropy, we have $D(\hat{p}_x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

Choosing α? Calibration curve technique

- $f(\alpha) =$ posterior expected log likelihood (fit to data).
- $g(\alpha) =$ posterior expected model complexity (effective complexity).
- $(g(\alpha), f(\alpha))$ traces out a curve in $\mathbb{R}^2$.
- Choose a point on this curve with good fit and low complexity.

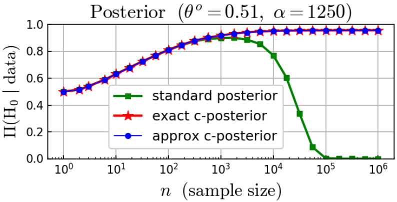


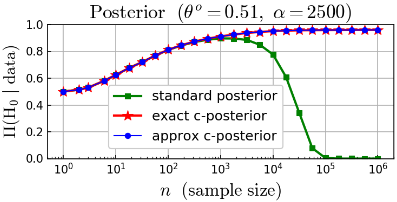




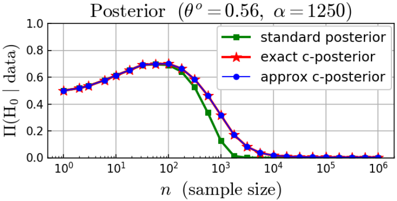
Toy example: Bernoulli trials

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 *a priori* 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.



Mixture models

- Model: $X_1, \ldots, X_n | w, \varphi$ i.i.d. $\sim \sum_{i=1}^{K} w_i f_{\varphi_i}(x)$
- Prior: $w \sim \mathrm{Dirichlet}(\gamma_1, \ldots, \gamma_K)$ and $\varphi_1, \ldots, \varphi_K \overset{\text{iid}}{\sim} H$.
- Relative entropy c-posterior is approximated by the power posterior,

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

  where $\zeta_n = \alpha/(\alpha + n)$.
- Could use Antoniano-Villalobos and Walker (2013) algorithm or RJMCMC (Green, 1995).
- We found a simple approximate algorithm that works well.

Algorithm: Conditional coarsening for mixtures

Same as standard data augmentation algorithm, except updates to w and φ use power likelihood.

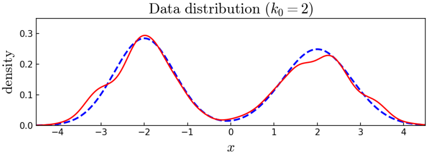
- Input: $x_1, \ldots, x_n$.
- Output: Samples of w, φ, and component assignments $z_1, \ldots, z_n$.

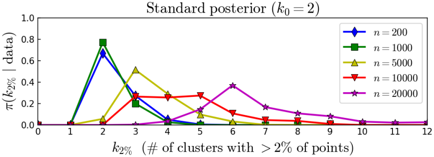
- For iteration $t = 1, \ldots, T$:
  1. For $j = 1, \ldots, n$: sample $z_j \sim \mathrm{Categorical}(\widetilde{w})$ where $\widetilde{w}_i \propto w_i f_{\varphi_i}(x_j)$.
  2. Sample $w \sim \mathrm{Dirichlet}(\widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_K)$ where $\widetilde{\gamma}_i = \gamma_i + \zeta_n \sum_{j=1}^n 1(z_j = i)$.
  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.

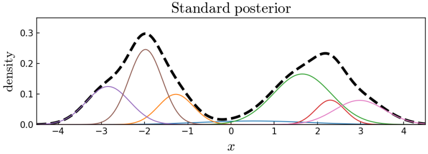
Algorithm: Conditional coarsening for mixtures

- Scales well to large datasets.
- Easy to implement.
- Yields results similar to (but not exactly the same as) the power posterior.

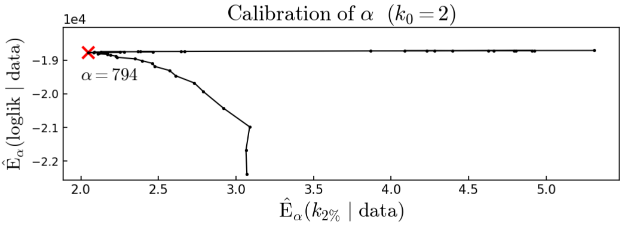
Example: Perturbed mixture of Gaussians


Data distribution ($k_0 = 2$)


Standard posterior ($k_0 = 2$)

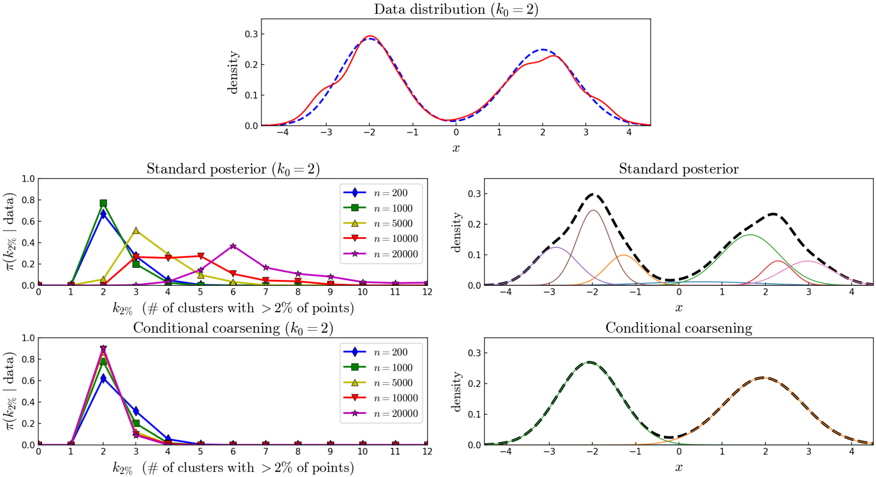

Standard posterior

Calibration curve for perturbed mixture of Gaussians

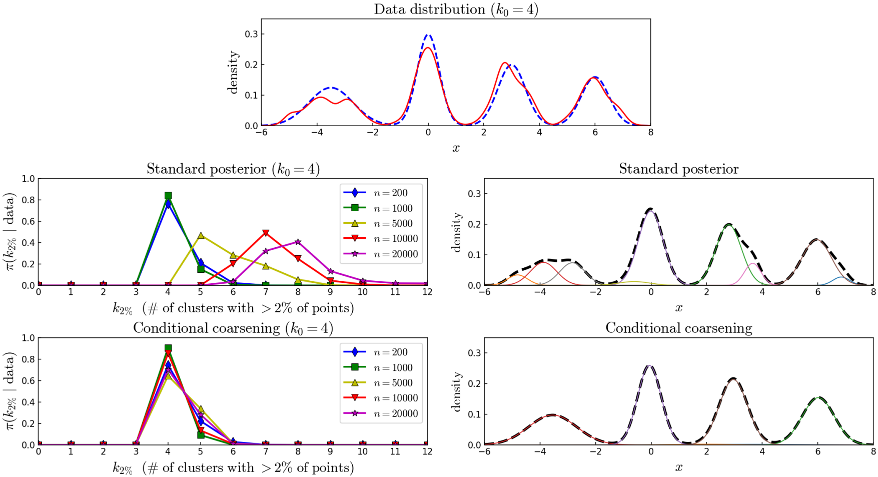


- $f(\alpha) =$ posterior expected log likelihood (fit to data).
- $g(\alpha) =$ posterior expected model complexity (effective complexity).
- $(g(\alpha), f(\alpha))$ traces out a curve in $\mathbb{R}^2$.
- Choose a point on this curve with good fit and low complexity.
- Suggests choosing $\alpha = 800$.

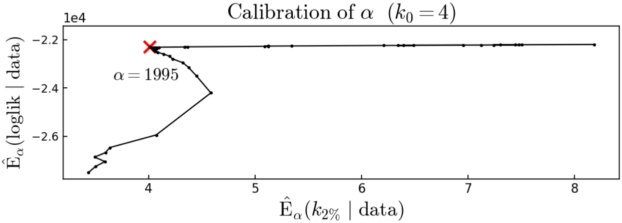
Example: Perturbed mixture of Gaussians



Example: Perturbed mixture of Gaussians



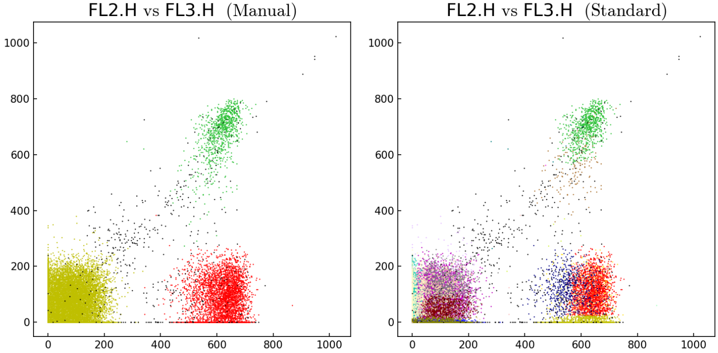
Calibration curve for perturbed mixture of Gaussians



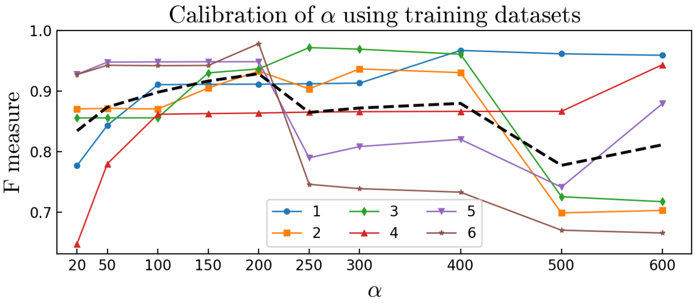
- $f(\alpha) =$ posterior expected log likelihood (fit to data).
- $g(\alpha) =$ posterior expected model complexity (effective complexity).
- $(g(\alpha), f(\alpha))$ traces out a curve in $\mathbb{R}^2$.
- Choose a point on this curve with good fit and low complexity.
- Suggests choosing $\alpha = 2000$.

Application: Flow cytometry clustering

- Each sample has 3 to 20-dim measurements on 10K's of cells.
- Manual clustering is time-consuming and subjective.
- Multivariate Gaussian mixture yields too many clusters.
- Example: GvHD data from FLOWCAP-I.



Calibration for flow cytometry

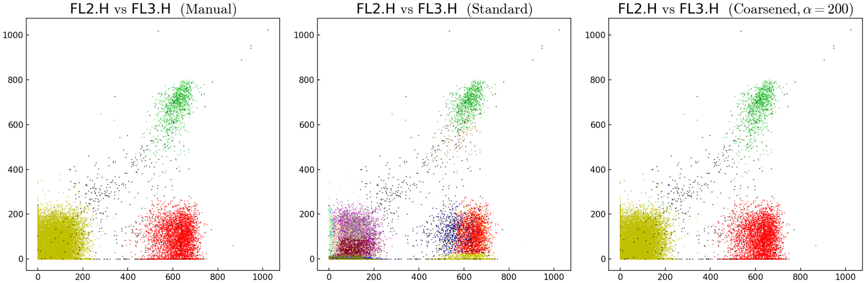


- Calibrate α using performance on GvHD datasets 1-6 for “training”.
- Best performance is at $\alpha = 200$ on training datasets.
- Use F-measure to quantify similarity of partitions $\mathcal{A}$ and $\mathcal{B}$:

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

Results: Flow cytometry clustering

Clustering on test datasets closely matches manual ground truth.



FL2.H vs FL3.H (Manual)

FL2.H vs FL3.H (Standard)

FL2.H vs FL3.H $(\text{Coarsened}, \alpha = 200)$

Results: Flow cytometry clustering

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

- Clustering on test datasets closely matches manual ground truth.
- Use F-measure to quantify similarity of partitions $\mathcal{A}$ and $\mathcal{B}$:

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

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,

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

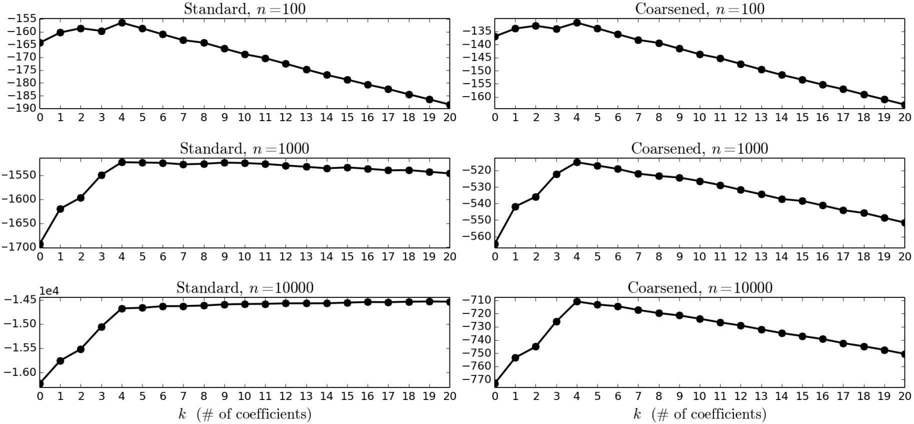
  where $\zeta_n = \alpha/(\alpha + n)$.
- This can be computed analytically, since $\theta|k$ has been given a conjugate prior.

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

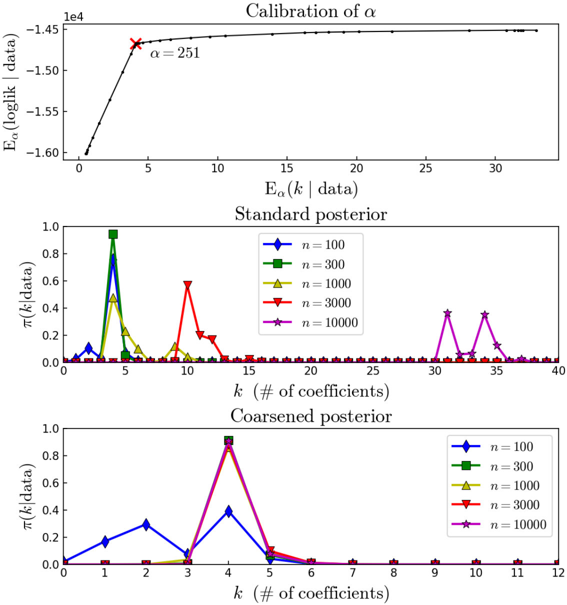
$$x_t = \tfrac{1}{4}(x_{t-1} + x_{t-2} - x_{t-3} + x_{t-4}) + \varepsilon_t + \tfrac{1}{2}\sin t$$

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Calibration curve suggests $\alpha = 250$.

Log of marginal likelihood



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



Outline

1. Motivation

2. Coarsened posterior, Power posterior

3. Examples
   - Toy example: Bernoulli trials
   - Mixture models and clustering
   - Autoregressive models of unknown order

4. Theory

Theory

We establish three main theoretical results:

1. large-sample asymptotics of c-posteriors as $n \to \infty$,
2. small-sample behaviour of c-posteriors, and
3. 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: Large-sample asymptotics

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*

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

and in fact,

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

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

Theory: Small-sample behaviour

- When n is small, the c-posterior tends to be well-approximated by the standard posterior.
- To study this, we consider the limit as the distribution of R converges to 0, while holding n fixed.

Theorem

Under regularity conditions, there exists $c_\alpha \in (0, \infty)$, not depending on θ, such that

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

- In particular, since $\zeta_n \approx 1$ when $n \ll \alpha$, the power posterior is a good approximation to the relative entropy c-posterior in this regime.

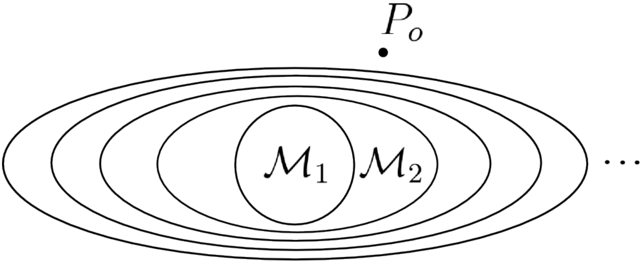
Theory: Lack of robustness of the standard posterior

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

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

  where $\doteq$ denotes agreement to first order in the exponent, i.e., $a_n \doteq b_n$ means $(1/n)\log(a_n/b_n) \to 0$.
- Due to the n in the exponent, even a slight change to P_o can dramatically change the posterior.

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 is 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,*

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

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.

Robust Bayesian inference via coarsening

Jeff Miller

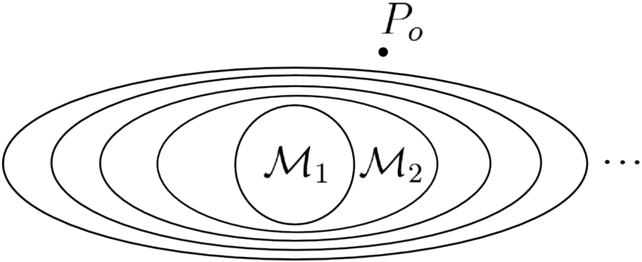
Joint work with David Dunson

Harvard University
Department of Biostatistics

Probability and Statistics Seminar, Boston University
March 15, 2018

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 $\mathcal{M}_k$.

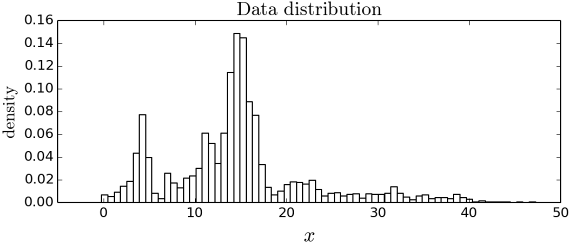


We seek an approach that tolerates models that are "close enough" to P_o.

Connection with ABC

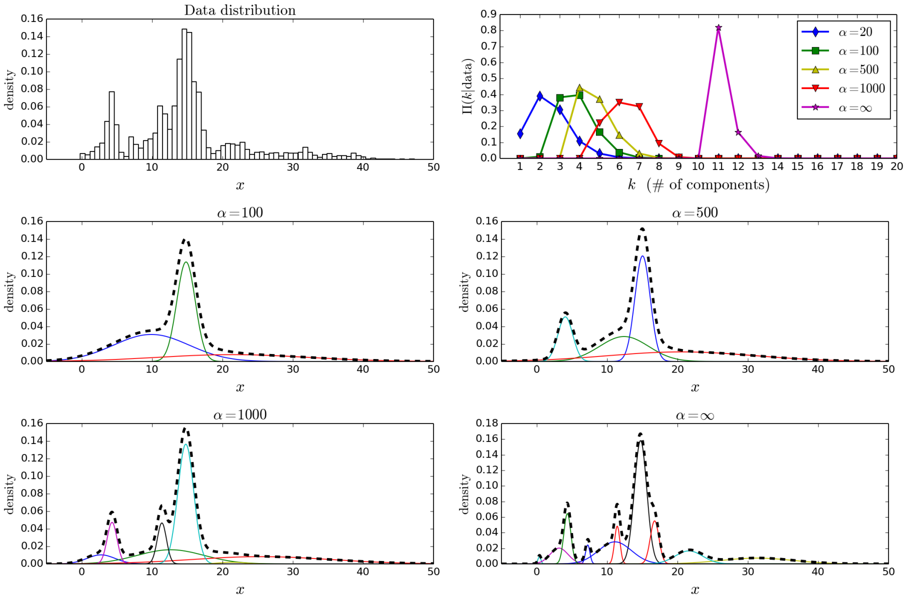
- The c-posterior $\pi\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big)$ is mathematically equivalent to the approximate posterior resulting from *approximate Bayesian computation* (ABC).
- Tavaré et al. (1997), Marjoram et al. (2003), Beaumont et al. (2002), Wilkinson (2013)
- However, there are some 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.

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



Example: Variable selection in linear regression

- Spike-and-slab model:

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

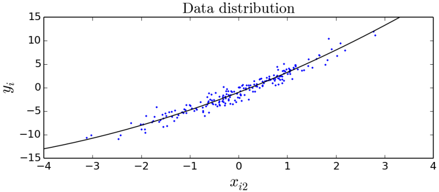
- For regression, a natural choice of statistical distance is conditional relative entropy. Again, this leads to a power posterior approximation to the c-posterior:

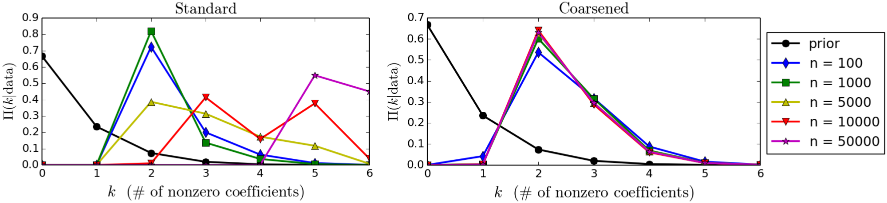
$$\pi\big(\beta, \sigma^2 \,\big|\, d_n(Y_{1:n}, y_{1:n}) < R\big) \mathrel{\underset{\sim}{\propto}} \pi(\beta, \sigma^2) \prod_{i=1}^{n} p(y_i | x_i, \beta, \sigma^2)^{\zeta_n}.$$

- 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

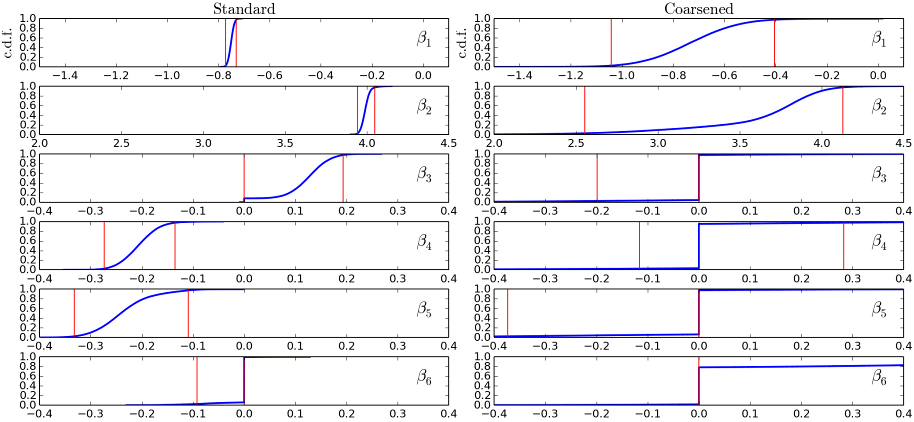
- Covariates: $x_{i1} = 1$ to accomodate constant offset, and $x_{i2}, \ldots, x_{i6}$ distributed according to a multivariate skew-normal distribution.
- $y_i = -1 + 4(x_{i2} + \frac{1}{16}x_{i2}^2) + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
- Set $\alpha = 50$,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

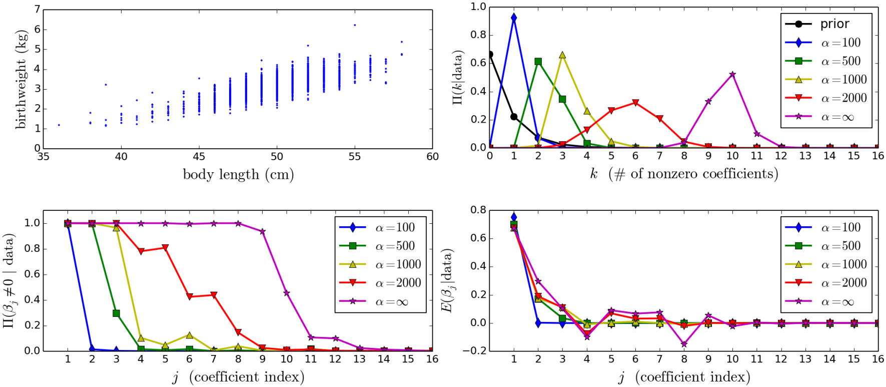
  $$\alpha \in \{100, 500, 1000, 2000, \infty\}$$

  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.