Several Interpretations of the Power Posterior

Jeff Miller

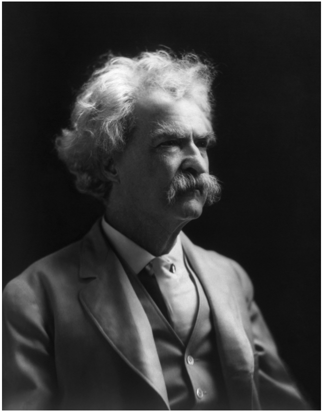
Joint work with David Dunson

Harvard University
T.H. Chan School of Public Health
Department of Biostatistics

BNP 11 // Paris // June 29, 2017

"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. Robust Bayes: different objectives $\Rightarrow$ different approaches

2. Robustness to perturbations

3. Interpretations of the power posterior

Outline

1. Robust Bayes: different objectives $\Rightarrow$ different approaches
2. Robustness to perturbations
3. Interpretations of the power posterior

Decision theoretic approaches to robust Bayes

- Standard Bayesian decision theory framework (Savage, 1954):

$$\min_{\text{action}} E(\text{loss}|\text{data}).$$

- Various minimax approaches are possible ...
- Robustness to the choice of prior (Berger 1984 and others):

$$\min_{\text{action}} \max_{\text{prior}\in\text{set}} E(\text{loss}|\text{data}).$$

- Robustness with respect to the posterior (Watson & Holmes 2016):

$$\min_{\text{action}} \max_{\text{posterior}\in\text{set}} E(\text{loss}|\text{data}).$$

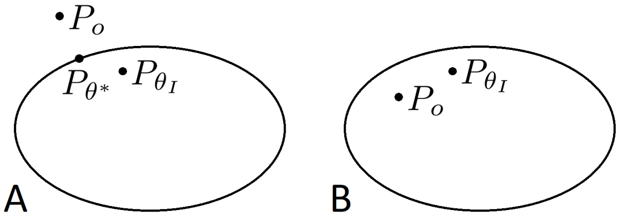
- Robustness to the choice of likelihood (anyone? seems interesting...):

$$\min_{\text{action}} \max_{\text{likelihood}\in\text{set}} E(\text{loss}|\text{data}).$$

- This talk focuses on robustness to misspecification of the likelihood.

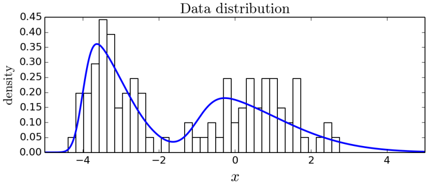
What do we mean by misspecification? Two scenarios

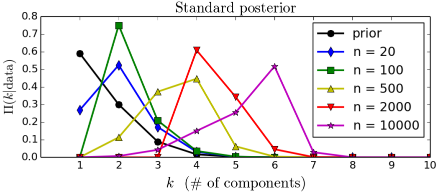
- Notation:
  - P_o = distribution of the observed data
  - θ^* = pseudo-true parameter (nearest point in model to P_o)
  - θ_I = ideal parameter (the truth before perturbation)
  - We think of P_o as a perturbation of P_{θ_I}.
- Scenario A: P_o is not in the model class.
- Scenario B: P_o is in the model class, but $P_o \neq P_{\theta_I}$.

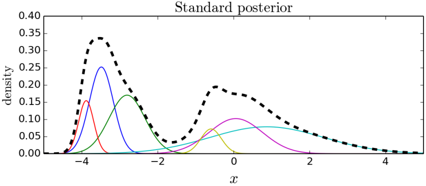


- If there is no perturbation, then $P_o = P_{\theta^*} = P_{\theta_I}$.

Example: Mixture models



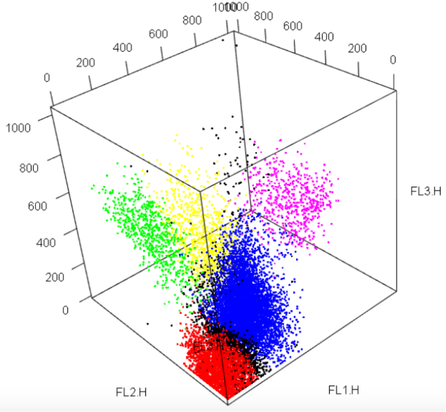




- P_{θ_I} is a two-component normal mixture, and P_o is a perturbation.
- The posterior introduces more and more components as n grows, in order to fit the data.
- P_o is approximable by a BNP mixture. . . but maybe we wanted θ_I!

Example: Flow cytometry

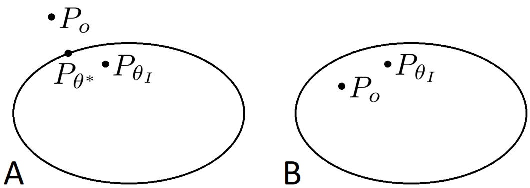
- Low-dim data with cell type clusters that are sort of Gaussian.
- Example: Graft versus Host Disease, $n = 13773$ blood cells, $d = 4$ flourescence signals, $K = 5$ manually labeled clusters of cell types.



(figure from Lee and McLachlan, Statistics and Computing, 2014)

What is the quantity of interest?

- The choice of method depends on the quantity of interest.
- Two main perspectives:
  1. *Fitting:* Model is a tool for approximating P_o.
     - ⋆ Want to predict future observations.
     - ⋆ Pseudo-true parameter θ^* is of interest.
  2. *Finding:* Model is an idealization of a true process.
     - ⋆ Want to recover unknown true parameters.
     - ⋆ Ideal parameter θ_I is of interest.



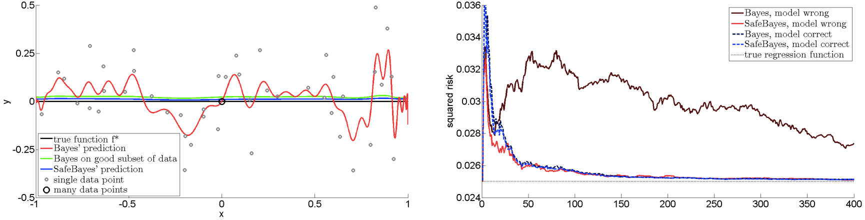
Perspective 1: Model is a tool for approximating P_o

- Pseudo-true parameter θ^* is of interest.
- Common when doing prediction using classification or regression.
- Examples:
  - Will person X get disease Y?
  - Will person X buy product Y?
  - How long will this person live?
  - What sentence was spoken in this recording?
  - What object is in this image?
  - Where are the tumors in this image?
  - What behavior is being exhibited by the mouse in this video?
  - Hot dog or no hot dog?
  - etc., etc., etc.

Issues with using standard posterior to infer θ^*

The posterior concentrates at θ^* (under regularity conditions), but ...

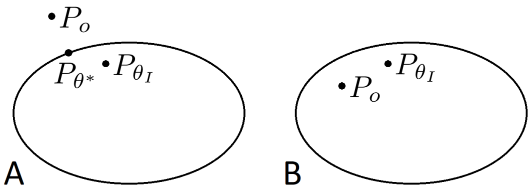
- Miscalibrated: credible sets do not have correct coverage
  - Kleijn & van der Vaart (2012)
  - Can recalibrate using sandwich covariance
- Slow concentration at the model containing θ^* can occur, leading to poor prediction performance
  - Grünwald & van Ommen (2014)
  - Can fix this using a power posterior $\propto p(x|\theta)^\zeta p(\theta)$ for certain $\zeta \in (0, 1)$



(figures from Grünwald & van Ommen, 2014)

Perspective 2: Model is an idealization of a true process

- Model is interpretable scientifically, but not exactly right of course.
- Ideal parameter θ_I is of interest.
- Data is from P_o, which we think of as a perturbation of P_{θ_I}.
- The objective is to understand — not to fit.
- This perspective is ubiquitous in science & medicine.



Perspective 2: Model is an idealization of a true process

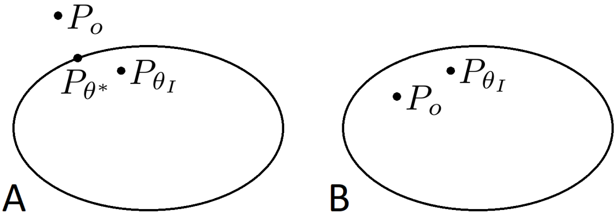
- Examples:
  - Phylogenetics
    - What is the evolutionary tree relating a given set of organisms?
  - Ecology
    - What factors affect which species live in which habitats?
  - Epidemiology
    - Does exposure X cause disease Y?
  - Cancer
    - What mutations occurred, and in what order?
  - Genomics / Genetics
    - Which genes are involved in causing disease Y?
  - Infectious diseases
    - How do infectious diseases spread?

Issues with using standard posterior to infer θ_I

- Lack of robustness
  - Small perturbations from P_{θ_I} can lead to large changes in the posterior. (e.g., mixture example)

- Miscalibration — too concentrated
  - If $P_o \neq P_{\theta_I}$, the posterior doesn't properly quantify uncertainty in θ_I.

"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."

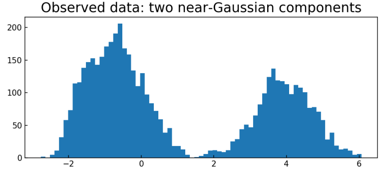


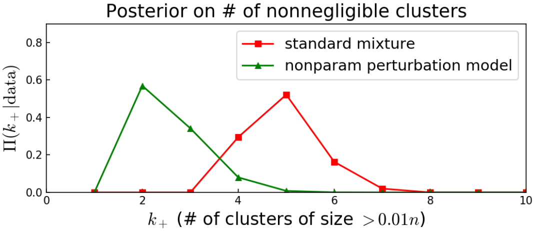
Outline

A BNP way to deal with perturbations

- Model $P_o|\theta_I$ using BNP.
  - Let's call this a NonParametric Perturbation (NPP) model
- Example: Perturbation of a finite mixture

$$\theta_I \sim \text{prior on finite mixtures}$$
$$P_o|\theta_I \sim \text{DP mixture with base measure } P_{\theta_I}$$
$$X_1, \ldots, X_n | P_o \sim P_o$$





A BNP way to deal with this

- Example (continued): Perturbation of a finite mixture.

  More detailed model description

  $\pi \sim \mathrm{Dirichlet}(\gamma_1, \ldots, \gamma_K)$
  $\mu_1, \ldots, \mu_K \sim \mathcal{N}(\mu_0, \sigma_0^2)$
  $\sigma_1^2, \ldots, \sigma_K^2 \sim \mathrm{InvGamma}(a_0, b_0)$
  $G|\pi, \mu, \sigma^2 \sim \mathrm{DP}\big(\alpha, \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \sigma_k^2)\big)$
  $X_1, \ldots, X_n|G \sim \int \mathcal{N}(x|y, s^2) dG(y)$

- Disadvantages:
  - More computationally burdensome
    - Have to introduce a bunch of auxiliary variables
  - More complicated
    - Scientists & doctors prefer methods they can understand
- Is there a simpler way to handle small perturbations?

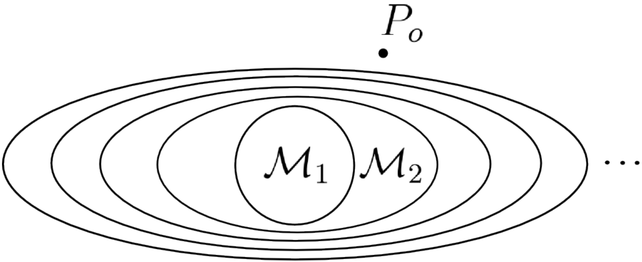
Lack of robustness of the standard posterior

- The standard posterior is not robust, especially for model inference. Why? Very roughly, if $x_i \sim P_o$ then when n is large,

$$p(\theta) \prod_{i=1}^{n} p_\theta(x_i) \mathrel{\underset{\sim}{\propto}} \exp(-nD(p_o \| p_\theta))p(\theta).$$

where $\mathrel{\underset{\sim}{\propto}}$ denotes approximate proportionality.

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

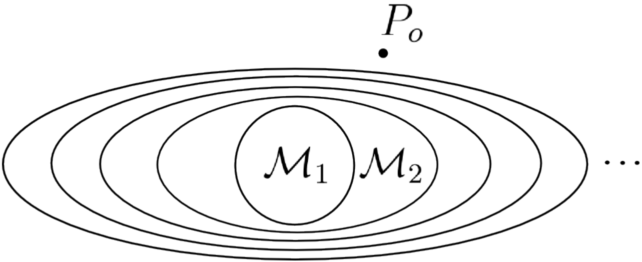


Intuition for how using a power posterior helps

- Raising the likelihood to a power $\zeta_n \in (0, 1)$, we get (very roughly)

$$p(\theta) \prod_{i=1}^{n} p_\theta(x_i)^{\zeta_n} \mathrel{\underset{\sim}{\propto}} \exp(-n\zeta_n D(p_o \| p_\theta)) p(\theta).$$

- Suppose $n\zeta_n \to \alpha$ and $D(p_o \| p_\theta)$ is close to $D(p_{\theta_I} \| p_\theta)$ as a function of θ.
- Then the power posterior given data from P_o will be close to the power posterior given data from P_{θ_I}, even as $n \to \infty$.

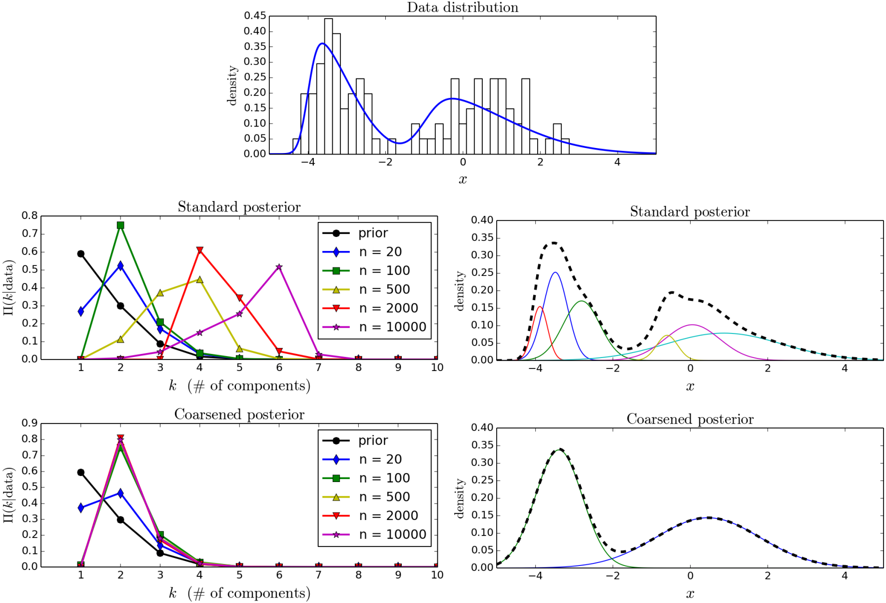


Outline

Interpretation 1: Changing the sample size

- The power posterior is only as concentrated as if we had $n\zeta_n$ samples.
- $\Rightarrow$ Can be viewed as changing n to $n\zeta_n$, in this sense.

Gaussian mixture applied to skew-normal mixture data



Interpretation 2: Balancing fit and model complexity

- By the Laplace approximation (under regularity conditions),

$$\log \int p(x_{1:n}|\theta_k)^{\zeta_n} p(\theta_k|k) d\theta_k \approx n\zeta_n \ell_n(k) - \tfrac{1}{2} D_k \log n + c_k$$

where D_k is the dimension of θ_k and

$$\ell_n(k) = \frac{1}{n} \log p(x_{1:n}|\hat{\theta}_k) \longrightarrow -D(p_o \| p_{\theta_k^*}) + \int p_o \log p_o.$$

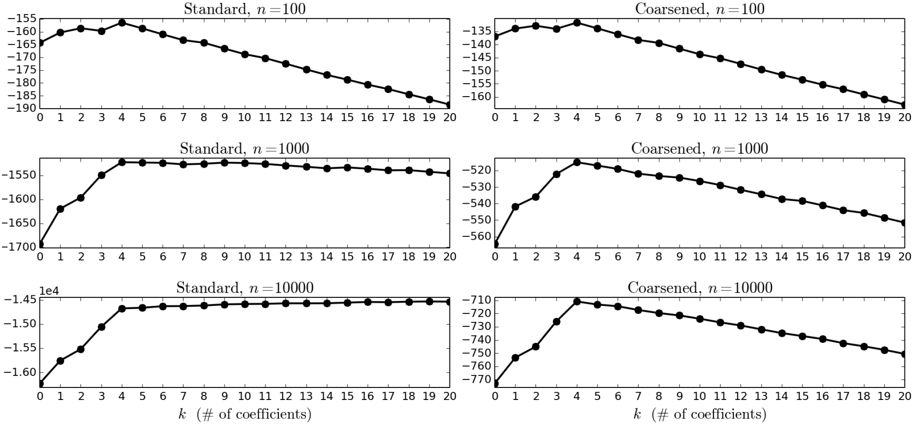
- $-\frac{1}{2} D_k \log n$ penalizes model complexity
- $n\zeta_n \ell_n(k)$ penalizes poor model fit to the data
- ζ_n allows one to balance these two penalties

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

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

where $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Choose $\zeta_n = \alpha/(\alpha+n)$ where $\alpha = 500$.

Log marginal likelihood vs model complexity k



Interpretation 3: Approximation to coarsened posterior

- Instead of the standard posterior $p(\theta \mid X_{1:n} = x_{1:n})$, M. & Dunson (2016) proposed the "coarsened posterior" (c-posterior)

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

  to obtain robustness to perturbations.
- Here, $d_n(X_{1:n}, x_{1:n}) \geq 0$ is a user-specified measure of the discrepancy between the empirical distributions $\hat{P}_{X_{1:n}}$ and $\hat{P}_{x_{1:n}}$.

Interpretation 3: Approximation to coarsened posterior

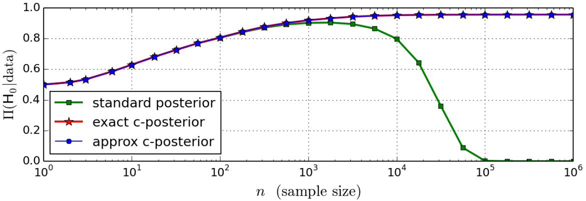
- Suppose $d_n(X_{1:n}, x_{1:n})$ is a consistent estimator of $D(p_o \| p_\theta)$ when $X_i \overset{\text{iid}}{\sim} p_\theta$ and $x_i \overset{\text{iid}}{\sim} p_o$.
- If $R \sim \mathrm{Exp}(\alpha)$ then we have the approximation

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

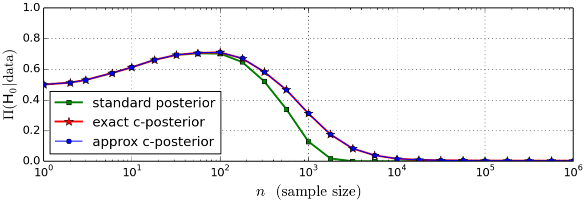
  where $\zeta_n = \alpha/(\alpha + n)$.
- This approximation is good when either $n \gg \alpha$ or $n \ll \alpha$, under mild conditions.

Toy example: Hypothesis testing with Bernoulli trials

Suppose $P_{\theta_I} = \text{Bernoulli}(0.5)$ and $P_o = \text{Bernoulli}(0.51)$. Consider $\text{H}_0 : \theta = 1/2$ versus $\text{H}_1 : \theta \neq 1/2$. Pick α to tolerate perturbations from θ_I of magnitude 0.02.



If $P_o = \text{Bernoulli}(0.56)$, the perturbation is significantly larger than our chosen tolerance. In both cases, the power posterior closely approximates the c-posterior.



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*

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

and in fact,

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

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.

Interpretation 4: Approximation to convolving the model

- The c-posterior can be expressed as:

$$
\begin{aligned}
p\big(\theta \mid d_n(X_{1:n}, x_{1:n}) < R\big) &\propto p(\theta)\, \mathbb{P}\big(d_n(X_{1:n}, x_{1:n}) < R \mid \theta\big) \\
&= p(\theta) \int G(d_n(x'_{1:n}, x_{1:n})) dP_\theta^n(x'_{1:n}),
\end{aligned}
$$

  where $G(r) = \mathbb{P}(R > r)$, e.g., if $R \sim \mathrm{Exp}(\alpha)$ then $G(r) = e^{-\alpha r}$.
- This integral can be viewed as a convolution of the model distribution P_θ^n with the "kernel" $G(d_n(x'_{1:n}, x_{1:n}))$.
- In cases where $G(d_n(x'_{1:n}, x_{1:n}))$ defines a distribution on $x_{1:n}$ given $x'_{1:n}$, the c-posterior is equivalent to integrating out this error distribution. However, even then, it will not necessarily be projective.

Other uses of power posteriors

- improving model selection & prediction performance under misspecification (Grünwald and van Ommen, 2014)
- discounting historical data (Ibrahim and Chen, 2000)
- obtaining consistency in BNP models (Walker & Hjort, 2001)
- marginal likelihood approximation (Friel and Pettitt, 2008)
- objective Bayesian model selection (O'Hagan, 1995)
- improved MCMC mixing (Geyer, 1991)

Several Interpretations of the Power Posterior

Jeff Miller

Joint work with David Dunson

Harvard University
T.H. Chan School of Public Health
Department of Biostatistics

BNP 11 // Paris // June 29, 2017