Robust Bayesian inference via coarsening

Jeff Miller

Joint work with David Dunson

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

MIT Machine Learning Colloquium // April 26, 2017

Outline

Outline

1. HSPH, Biostatistics, and Current projects
2. Coarsened posterior
3. Examples
   - Mixture models with an unknown number of components
   - Variable selection in linear regression
4. Theory

Longwood Medical Area

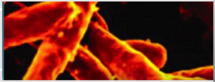
You want medical, we got medical: Beth Israel, Brigham & Women's, Dana Farber, Children's, Harvard Medical School, HSPH.

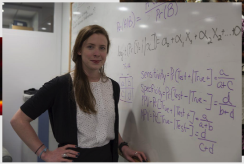
Harvard T.H. Chan School of Public Health

Academic Departments

- Biostatistics
- Environmental Health
- Epidemiology
- Genetics and Complex Diseases
- Global Health and Population
- Health Policy and Management
- Immunology and Infectious Diseases
- Nutrition
- Social and Behavioral Sciences

© Alex Hofford

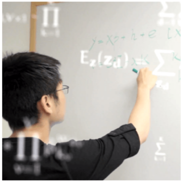




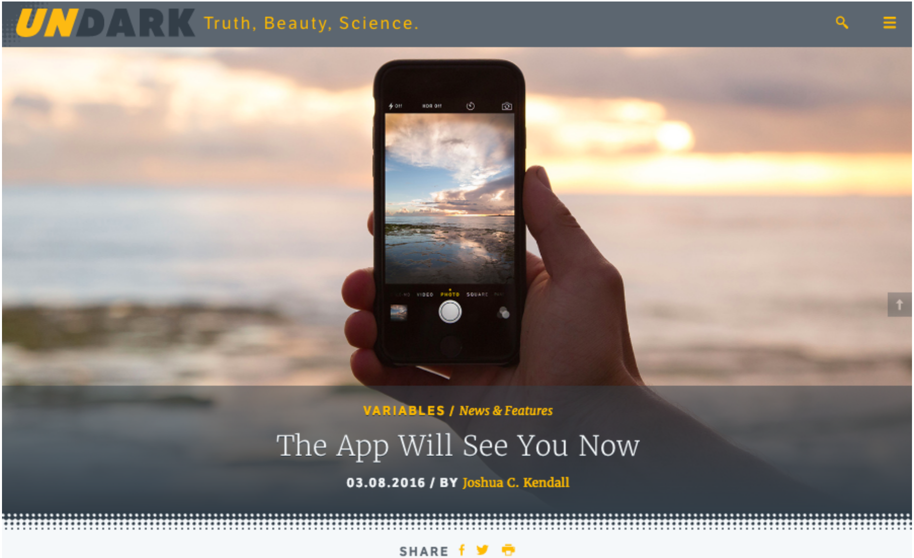
Biostatistics department — broad range of topics

- Genomics of complex diseases
- Environmental statistics
- Causal inference
- Cancer genomics
- Neurostatistics
- HIV, infectious diseases
- Epidemiology
- Clinical trials



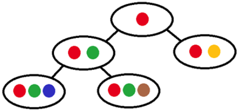


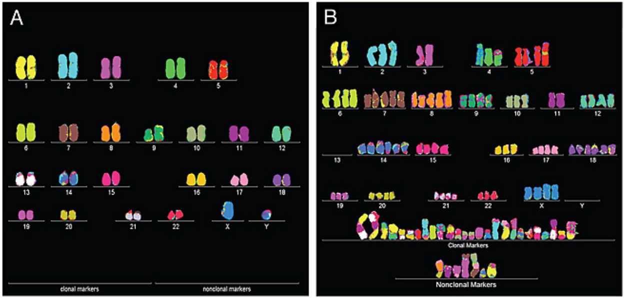
Digital phenotyping (J.P. Onnela)



Cancer phylogenetic inference

- Cancer evolves into multiple populations within a patient.
- Problem is to deconvolve populations and recover phylogenetic tree.
- Collab. with Scott Carter (DFCI), using whole-exome/whole-genome.
- Using hybrid of Bayes & frequentist — VB mixtures, hyp. tests, . . .

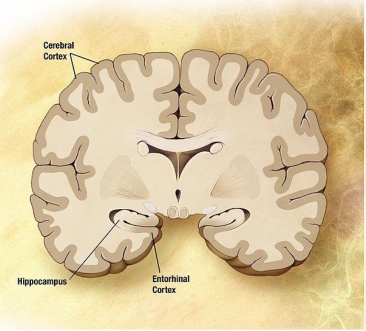


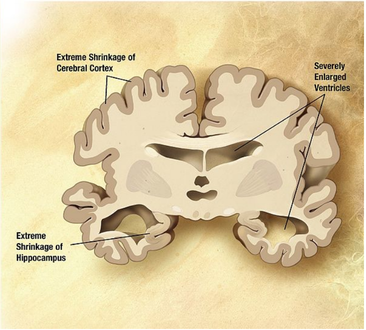


http://news.berkeley.edu/2011/07/26/are-cancers-newly-evolved-species/

Studying Alzheimer's with whole-genome sequences

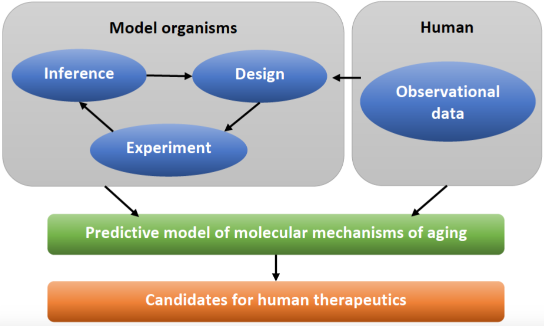
- Collaboration with Rudy Tanzi (MGH), Christoph Lange (HSPH).
- 1971 whole-genome seqs from 558 families (NIMH+NIA).
- By using family relations, can condition away many confounders.
- Using Generalized Higher Criticism for powerful GWAS tests.
- Working on moving beyond traditional GWAS . . .





Inference, Design of experiments, and Experimentation in an Automated Loop (IDEAL) for aging research

- Recently-developed automated parallel experimentation devices.
- e.g., Fontana lab at HMS built "Lifespan machine" performing experiments on 10,000s of C. elegans worms simultaneously.
- Need optimal experimental design methods to fully exploit.



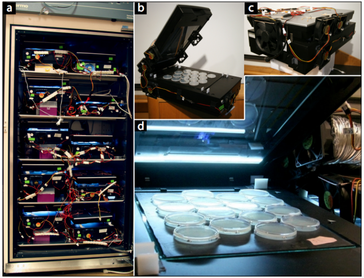


Image from Stroustrup et al., Nature Methods, 2013.

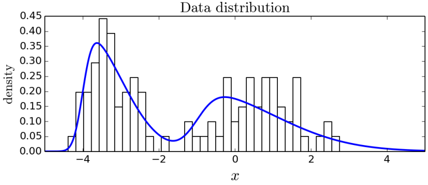
Outline

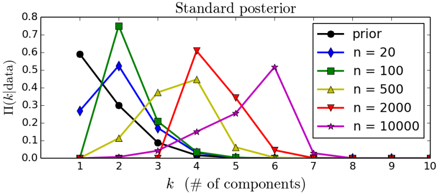
1. HSPH, Biostatistics, and Current projects
2. Coarsened posterior
3. Examples
   - Mixture models with an unknown number of components
   - Variable selection in linear regression
4. Theory

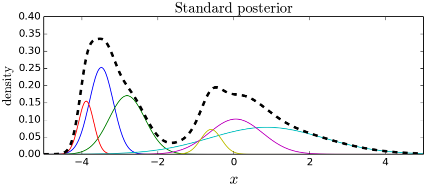
Motivation

- 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."
- Ideally, one would use a completely correct model, but this is often impractical.

Example: Mixture models

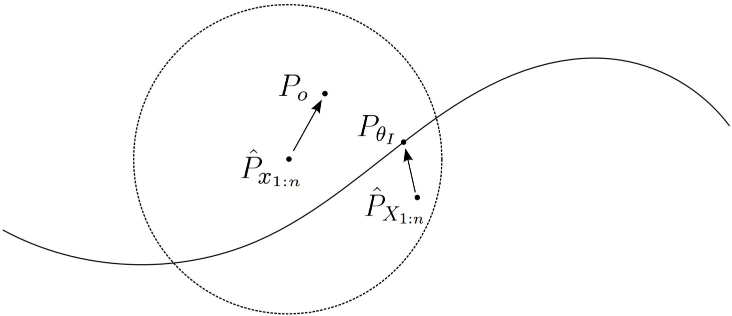






- Mixtures are often used for clustering.
- But if the data distribution is not exactly a mixture from the assumed family, the posterior will often introduce more and more clusters as n grows, in order to fit the data.
- As a result, the interpretability of the clusters may break down.

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 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.

Relative entropy c-posteriors

- There are many possible choices of statistical distance . . .

  e.g., KS, Wasserstein, maximum mean discrepancy, various divergences

  . . . but relative entropy (KL divergence) works out exceptionally nicely.
- Define $d_n(X_{1:n}, x_{1:n})$ to be a consistent estimator of $D(p_o \| p_\theta)$ when $X_i \overset{\text{iid}}{\sim} p_\theta$ and $x_i \overset{\text{iid}}{\sim} p_o$. (Recall: $D(p_o \| p_\theta) = \int p_o(x) \log \frac{p_o(x)}{p_\theta(x)} dx$.)
- 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 $\zeta_n = \alpha/(\alpha + n)$. This approximation is good when either $n \gg \alpha$ or $n \ll \alpha$, under mild conditions.
- 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

Recent work on Bayesian robustness

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

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

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)
- Sometimes, this is done to ensure appropriate concentration at the minimal KL point when the model is misspecified.
  - Royall and Tsou (2003)
  - Grünwald and van Ommen (2014)
- However, the form of power we use, and its theoretical justification, seem novel.

Interpretation of power posterior

- Using the power posterior $\propto \pi(\theta) \prod_{i=1}^{n} p_\theta(x_i)^\zeta$ corresponds to adjusting the sample size from n to $n\zeta$, in the sense that the posterior will only be as concentrated as if there were $n\zeta$ samples.
- Thus, by setting $\zeta = \alpha/(\alpha + n)$, one makes the power posterior tolerant (asymptotically) of all θ's for which a sample of size α could plausibly have come from P_θ.

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.

Outline

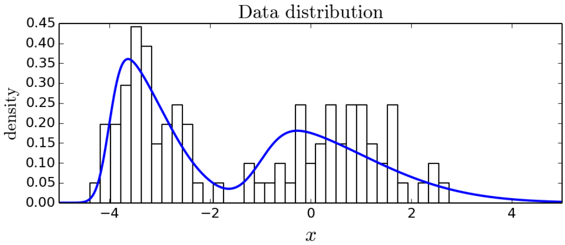
Example: Gaussian mixture with a prior on k

- 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,

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

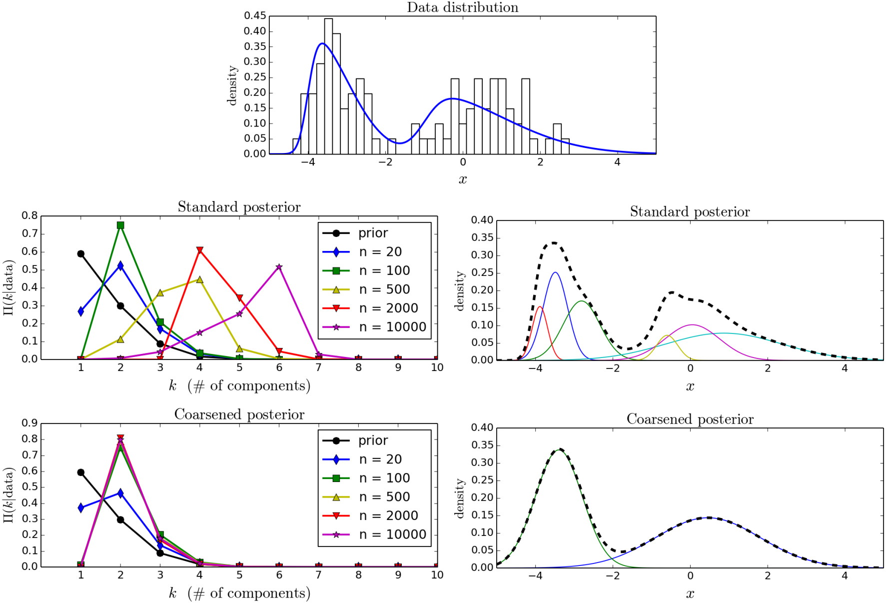
  where $\zeta_n = \alpha/(\alpha + n)$.
- Could use Antoniano-Villalobos and Walker (2013) algorithm or RJMCMC (Green, 1995). For simplicity, we reparametrize in a way that allows the use of plain-vanilla Metropolis–Hastings.

Gaussian mixture applied to skew-normal mixture data

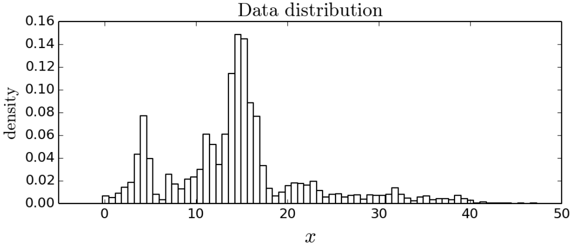


- Data: $x_1, \ldots, x_n$ i.i.d. $\sim \frac{1}{2}\mathcal{SN}(-4, 1, 5) + \frac{1}{2}\mathcal{SN}(-1, 2, 5)$, where $\mathcal{SN}(\xi, s, a)$ is the skew-normal distribution with location ξ, scale s, and shape a (Azzalini and Capitanio, 1999).
- Use strategy #2: Choose $\alpha = 100$, to be robust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

Gaussian mixture applied to skew-normal mixture data

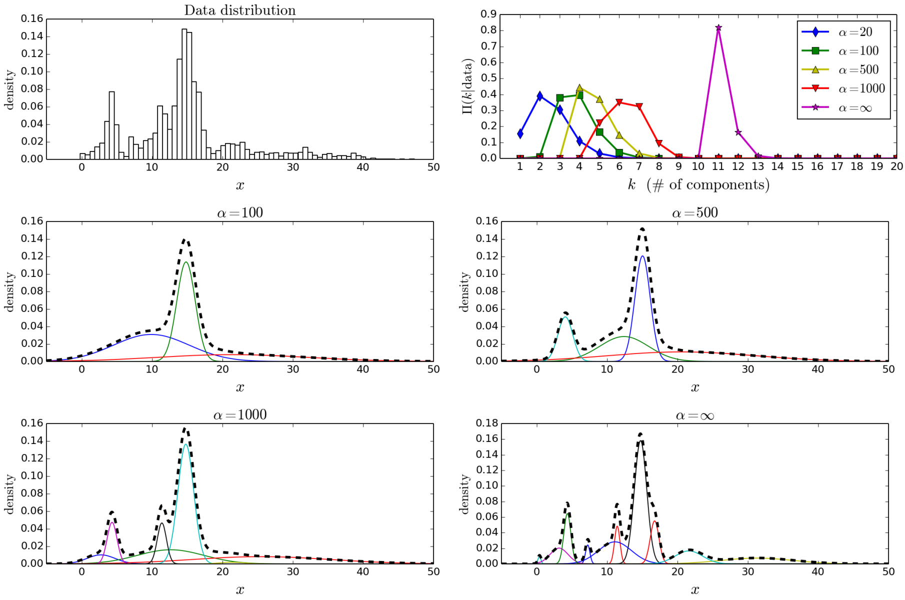


Velocities of galaxies in the Shapley supercluster



- 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



Example: Variable selection in linear regression

- Spike-and-slab model:

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

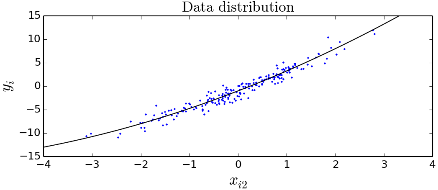
- For regression, a natural choice of statistical distance is conditional relative entropy. Again, this leads to a power posterior approximation to the c-posterior:

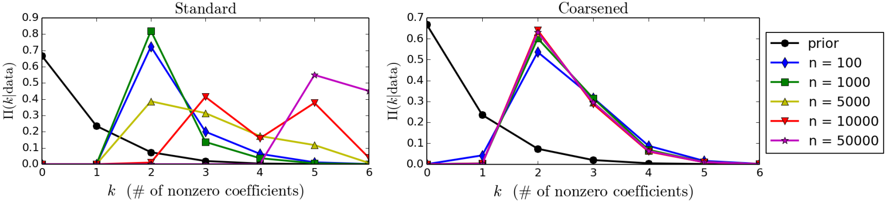
$$\pi\big(\beta, \sigma^2 \,\big|\, d_n(Y_{1:n}, y_{1:n}) < R\big) \mathrel{\underset{\sim}{\propto}} \pi(\beta, \sigma^2) \prod_{i=1}^{n} p(y_i | x_i, \beta, \sigma^2)^{\zeta_n}.$$

- 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

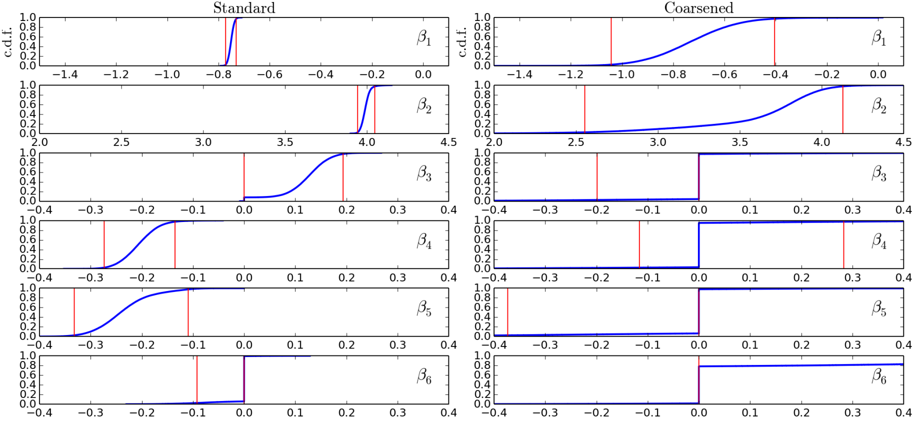
- Covariates: $x_{i1} = 1$ to accomodate constant offset, and $x_{i2}, \ldots, x_{i6}$ distributed according to a multivariate skew-normal distribution.
- $y_i = -1 + 4(x_{i2} + \frac{1}{16}x_{i2}^2) + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.
- Set $\alpha =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

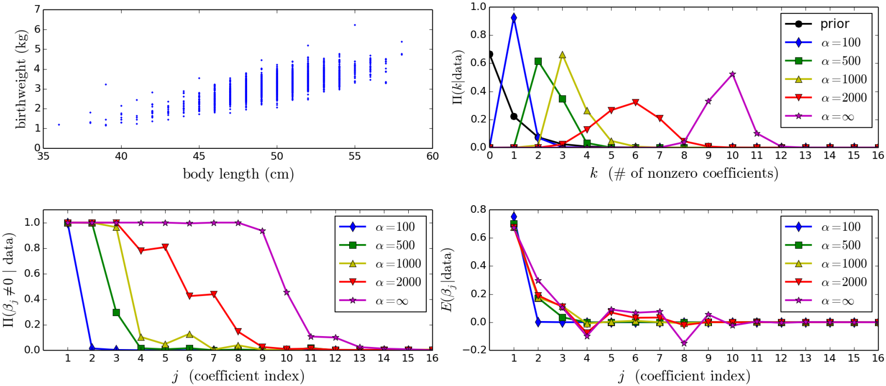
$$\alpha \in \{100, 500, 1000, 2000, \infty\}$$

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.

Outline

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,

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

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

Theory: 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.

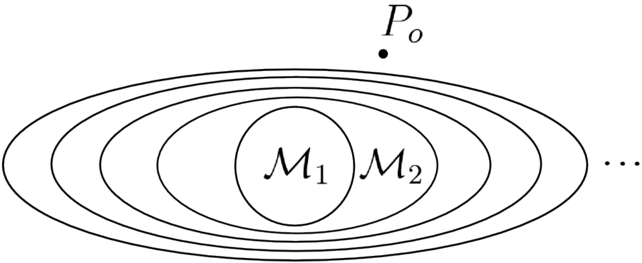
Theory: Lack of robustness of the standard posterior

- 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

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

  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,

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

Future work

- Can we choose α adaptively, to obtain consistency when the model is correct, and appropriate calibration otherwise?
  - (Well, yes, but can we do it in a computationally efficient way?)
- Looking at possible applications in causal inference.
- Develop inverse specification approach.

We have lots of biomedical data and challenging problems — if anyone is interested in collaborating let me know!

Thank you!