Inference using Partial Information

Jeff Miller

Harvard University
Department of Biostatistics

ICERM Probabilistic Scientific Computing workshop
June 8, 2017

Outline

1. Partial information: What? Why?
2. Need for modular inference framework
3. Cancer phylogenetic inference
4. Coarsening for robustness

What does it mean to use partial information?

What does it mean to use partial information?

Be ignorant.

What does it mean to use partial information?

Be ignorant.

In other words, ignore part of the data, or part of the model.

Why use partial info? Speed, simplicity, & robustness

- The Neyman–Scott problem is a very simple but nice example:
- Suppose $X_i, Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ indep. for $i = 1, \ldots, n$, and we want to infer σ^2, but the distribution of the μ's is completely unknown.
- Problem: MLE is inconsistent, and using the wrong prior on the μ's leads to inconsistency.
- Bayesian approach: Put a prior on the distribution of the μ's, e.g., use a Dirichlet process mixture and do inference with usual algorithms.
- Partial info approach: Let

  $$Z_i = X_i - Y_i \sim \mathcal{N}(0, 2\sigma^2)$$

  and use $p(z_1, \ldots, z_n | \sigma^2)$ to infer σ^2. Way easier!
- Partial model gives consistent and correctly calibrated Bayesian posterior on σ^2 — just slightly less concentrated.

More general example: Composite posterior

- Suppose we have a model $p(x|\theta)$ (where x is all of the data).
- We could do inference based on $p(s|t,\theta)$ for some statistics $s(x)$ and $t(x)$, i.e., ignore info in $p(t|\theta)$ and $p(x|s,t,\theta)$.
- Or, could combine and use $\prod_i p(s_i|t_i,\theta)$ for some $s_i(x)$ and $t_i(x)$.
  - This is Lindsay's composite likelihood.
- Composite MLE is

$$\hat{\theta}_n = \operatorname*{argmax}_\theta \prod_{i=1}^n p(s_i|t_i,\theta).$$

- Can define "composite posterior":

$$\pi_n(\theta) \propto p(\theta) \prod_{i=1}^n p(s_i|t_i,\theta).$$

  - When is this valid? i.e., correctly calibrated in a frequentist sense?

Composite posterior calibration

- Under regularity conditions, $\hat{\theta}_n$ is asymptotically normal:

$$\hat{\theta}_n \approx \mathcal{N}(\theta_0, A_n^{-1} C_n A_n^{-1})$$

when $X \sim p(x|\theta_0)$, where $g_i(x, \theta) = \nabla_\theta \log p(s_i(x) \mid t_i(x), \theta)$,

$$A_n = \sum_{i=1}^n \mathrm{Cov}\big(g_i(X, \theta_0)\big), \qquad C_n = \mathrm{Cov}\Big(\textstyle\sum_{i=1}^n g_i(X, \theta_0)\Big).$$

- Meanwhile, under regularity conditions, π_n is asymptotically normal:

$$\pi_n(\theta) \approx \mathcal{N}(\theta \mid \hat{\theta}_n, A_n^{-1}).$$

- When $g_1(X, \theta_0), \ldots, g_n(X, \theta_0)$ are uncorrelated, $A_n = C_n$.
- In this case, the composite posterior is well-calibrated in terms of frequentist coverage (asymptotically, at least).

Usage of partial information

- Frequentists use partial information all the time:
  - Composite likelihoods (partial likelihood, conditional likelihood, pseudo-likelihood, marginal likelihood, rank likelihood, etc.)
  - Generalized method of moments, Generalized estimating equations
  - Tests based on insufficient statistics (many methods here)
- But Bayesians try to avoid information loss.
  - Exceptions:
    - Using subsets of data for computational speed
    - Scattered usage of composite posteriors: Doksum & Lo (1990), Raftery, Madigan, & Volinsky (1996), Hoff (2007), Liu, Bayarri, & Berger (2009), Pauli, Racugno, & Ventura (2011).
  - Main issue is ensuring correct calibration of generalized posteriors.
  - In recent work, we have developed Bernstein–Von Mises results for generalized posteriors, to facilitate correct calibration.

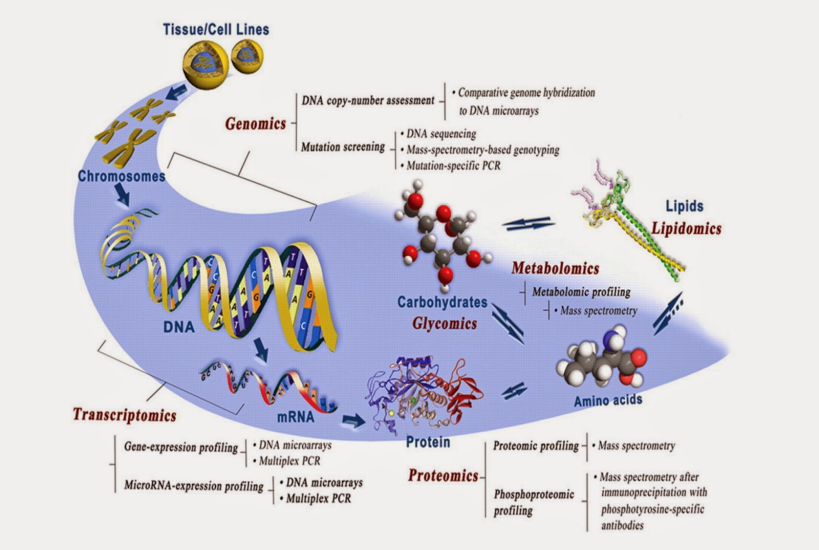
Outline

1. Partial information: What? Why?
2. Need for modular inference framework
3. Cancer phylogenetic inference
4. Coarsening for robustness

Need for modular inference framework

- Large complex biomedical data sets are currently analyzed by *ad hoc* combinations of tools, each of which uses partial info.
- We need a sound framework for combining tools in a modular way.

Diverse 'omics data types



from Wu et al. JDR 2011, 90:561-572

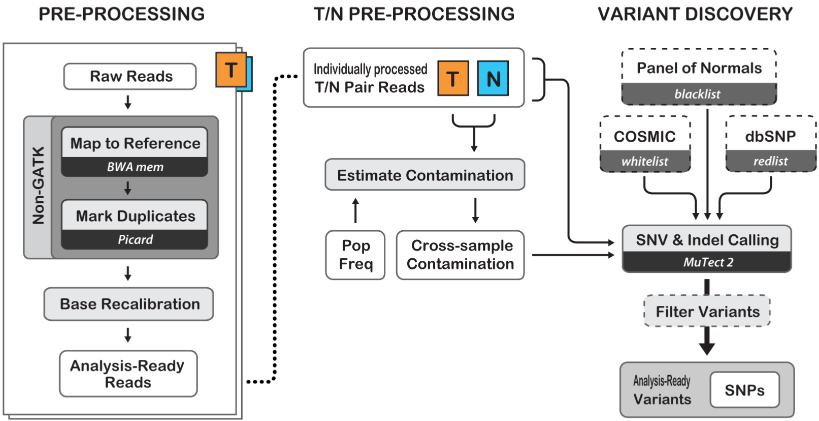
Motivation

- Biomedical data sets grow ever larger and more diverse.
- For example, the TOPMed program of the National Heart, Lung, and Blood Institute (NHLBI) is collecting:
  - whole genome, methylation, gene expression, proteome, metabolome
  - molecular, behavioral, imaging, environmental, and clinical data
  - for approximately 120,000 individuals
- Data collections like this will continue to grow in number and scale.

Challenge: Specialized methods are required

- These data are complex, requiring carefully tailored statistical and computational methods.
- Issues:
  - raw data very indirectly related to quantities of interest
  - selection effects, varying study designs (family, case-control, cohort)
  - missing data (e.g., 80-90% missing in single-cell DNA methylation)
  - batch/lab effects make it tricky to combine data sets
  - technical artifacts and biases in measurement technology
- As a result, many specialized tools have been developed, each of which solves a subproblem.
- These tools are combined into analysis "pipelines".

Example: Cancer genomics pipeline



Best Practices for Somatic SNVs and Indels in Whole Genomes and Exomes - BETA

from Broad Institute, Genome Analysis Toolkit (GATK) documentation

Example: Cancer genomics pipeline (continued)

. . . then:

- Indelocator – detect small insertions/deletions (indels)
- MutSig – prioritize mutations based on inferred selective advantage
- ContEst – contamination estimation and filtering
- HapSeg – estimate haplotype-specific copy ratios
- GISTIC – identify and filter germline chromosomal abnormalities
- Absolute – estimate purity, ploidy, and absolute copy numbers
- Manual inspection and analysis

- Many of these tools use statistical models and tests, but there is no overall coherent model.

Pros and cons of using partial info and then combining

- Cons:
  - Issues with uncertainty quantification
  - Loss of information
  - Potential biases, lack of coherency
- Pros:
  - Computational efficiency
  - Robustness to model misspecification
  - Reliable performance
  - Modularity, flexibility, and ease-of-use
  - Facilitates good software design

    Write programs that do one thing and do it well.
    Write programs to work together.
  - Division of labor (both in development and use)
- Ideally, we would use a single all-encompassing probabilistic model. But this is not practical for a variety of reasons.

Moral: We need a framework for modular inference

- Monolithic models are not well-suited for large complex data.
- The (inevitable?) alternative is to use modular methods based on partial information.
- Question: How to combine methods in a coherent way?
- We need a sound statistical framework for combining methods that each solve part of an inference problem.

Outline

1. Partial information: What? Why?

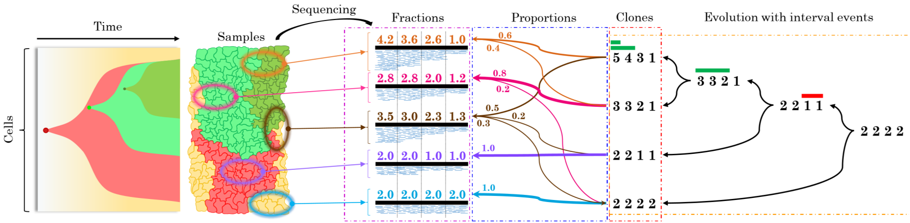
2. Need for modular inference framework

3. Cancer phylogenetic inference

4. Coarsening for robustness

Cancer phylogenetic inference (Joint work with Scott Carter)

- Cancer evolves into multiple populations within each person.
- Genome sequencing of tumor tissue samples is used for treatment.
- In bulk sequencing, each sample has cells from multiple populations.
- Goal: Infer the number of populations, their mutation profiles, and the phylogenetic tree.



from Zaccaria, Inferring Genomic Variants and their Evolution, 2017

Cancer phylogenetic inference

Parameters / latent variables:

- K = number of populations.
- Tree T on populations $k = 1, \ldots, K$.
- Copy numbers: q_{km} = # copies of segment m in a cell from pop k.
- Proportions: p_{sk} = proportion of cells in sample s from population k.

Model (leaving several things out, to simplify the description):

- Branching process model for T and K
- Markov process model for copy numbers Q
- Dirichlet priors for proportions P
- Data:

$$X = PQ + \varepsilon$$

  where $\varepsilon_{sm} \sim \mathcal{N}(0, \sigma^2_{sm})$.

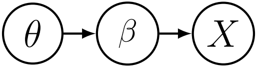
Cancer phylogenetic inference

Inference:

- MCMC and Variational Bayes do not work well (believe me, I tried!)
- Difficulty: Large combinatorial space with many local optima.
- We really care about the true tree – not just fitting the data.

New(?) idea: Method of sufficient parameters

- Idea: Temporarily ignore some dependencies among parameters.
- Consider a model $p(x|\theta)$ (where x is all of the data).
- Suppose $\beta = \beta(\theta)$ is such that $X \perp \theta \mid \beta(\theta)$.

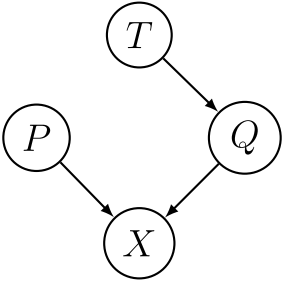


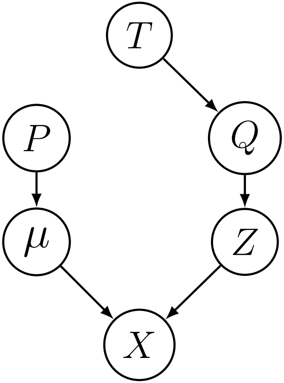
Method:

1. Infer β using $p(x|\beta)$.
    - **Ignore constraints** on β due to its definition as a function of θ.
    - Use a convenience prior on β (not the induced prior from $p(\theta)$).
2. Infer θ from β.
    - e.g., use $p(\theta|\beta)$.
3. Use 1 and 2 to construct an importance sampling (IS) distn for θ.
    - Use IS for posterior inference from the exact posterior $p(\theta|x)$.

Sufficient parameters for cancer phylo problem

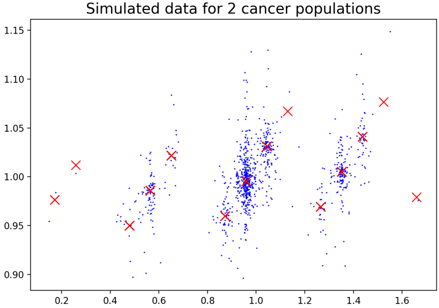
- Recall our data model: $X = PQ + \varepsilon$ where $\varepsilon_{sm} \sim \mathcal{N}(0, \sigma_{sm}^2)$.
- Given P, the columns of X are draws from a Gaussian mixture model with component means $\mu_i = P v_i \in \mathbb{R}^S$ for some $v_1, v_2, \ldots \in \mathbb{Z}^K$.
- We take $\beta = (\mu, Z)$ as our sufficient parameters, where $Z = (Z_1, \ldots, Z_M)$ is the component assignments, and $\theta = (T, P, Q)$.

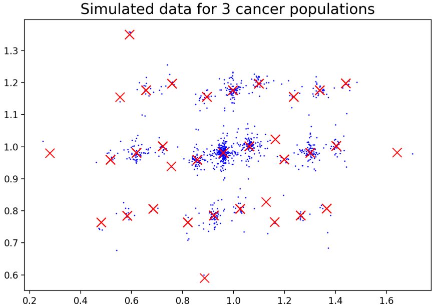




Sufficient parameters for cancer phylo problem

- Can infer the means μ and component assignments Z from X using a standard Gaussian mixture model algorithm.
  - The means form a lattice, but we ignore this constraint in this step.
  - More generally, we ignore the prior on (μ, Z) induced by (T, P, Q). Instead, we use Gaussian and Dirichlet-Categorical priors on μ and Z.
- We can then infer (T, P, Q) from (μ, Z) using other methods.


Simulated data for 2 cancer populations


Simulated data for 3 cancer populations

Demo

- True tree: $\tau = [0, 1, 1, 3, 3]$ where $\tau_i =$ parent of i.
- Ranked list of trees that are consistent with the data:

rank	tree	score
1:	[0,1,1,3,3]	0.305 (true)
2:	[0,3,1,3,1]	0.176
3:	[0,1,1,3,1]	0.000

- 97% of mutation profile correctly estimated.
- (This example uses point mutations – similar but slightly different.)

Demo

- True tree: $\tau = [0, 1, 2, 2, 3, 2, 4, 4]$ where τ_i = parent of i.
- Ranked list of trees that are consistent with the data:

rank	tree	score
1:	[0,1,2,2,3,2,4,4]	0.007525 (true)
2:	[0,1,2,2,3,4,2,4]	0.004130
3:	[0,1,2,2,3,7,2,4]	0.000260
4:	[0,1,2,2,3,7,4,2]	0.000260
5:	[0,1,2,2,3,7,4,4]	0.000260
6:	[0,1,2,2,3,4,4,2]	0.000007

- 92% of mutation profile correctly estimated.
- (This example uses point mutations – similar but slightly different.)

Outline

1 Partial information: What? Why?

2 Need for modular inference framework

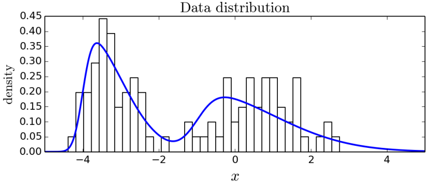
3 Cancer phylogenetic inference

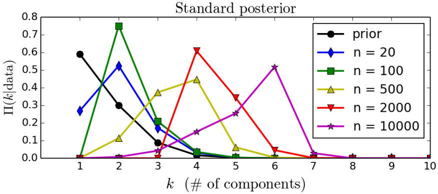
4 Coarsening for robustness

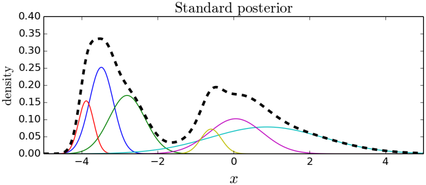
Motivation

- 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

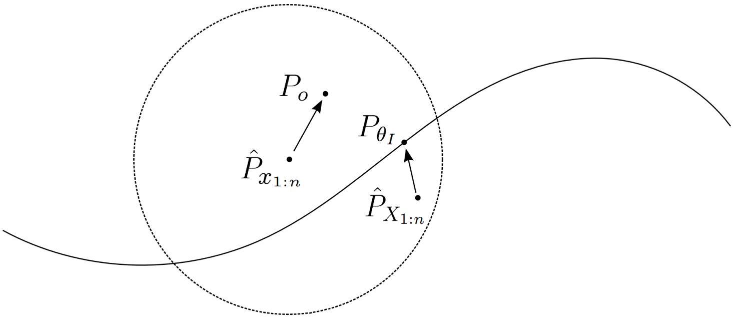






- Mixtures are often used for clustering.
- But if the data distribution is not exactly a mixture from the assumed family, the posterior will often introduce more and more clusters as n grows, in order to fit the data.
- As a result, the interpretability of the clusters may break down.

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

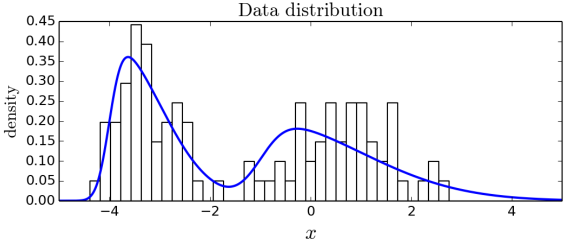
Example: Gaussian mixture with a prior on k

- 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,

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

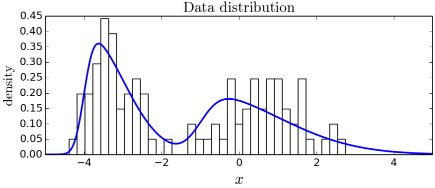
  where $\zeta_n = \alpha/(\alpha + n)$.
- Could use Antoniano-Villalobos and Walker (2013) algorithm or RJMCMC (Green, 1995). For simplicity, we reparametrize in a way that allows the use of plain-vanilla Metropolis–Hastings.

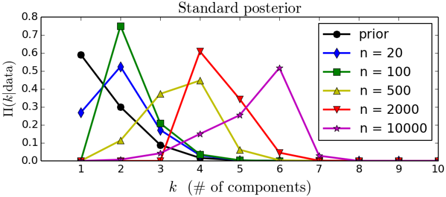
Gaussian mixture applied to skew-normal mixture data

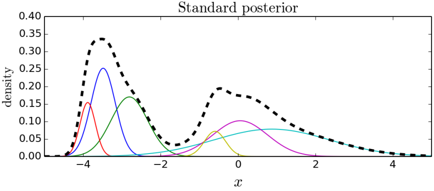


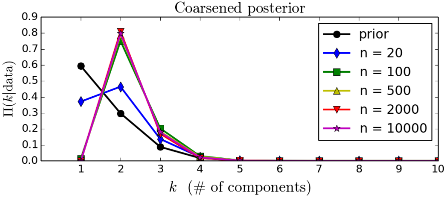
- Data: $x_1, \ldots, x_n$ i.i.d. $\sim \frac{1}{2}\mathcal{SN}(-4,1,5) + \frac{1}{2}\mathcal{SN}(-1,2,5)$, where $\mathcal{SN}(\xi, s, a)$ is the skew-normal distribution with location ξ, scale s, and shape a (Azzalini and Capitanio, 1999).
- Choose $\alpha = 100$, to be robust to perturbations to P_o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

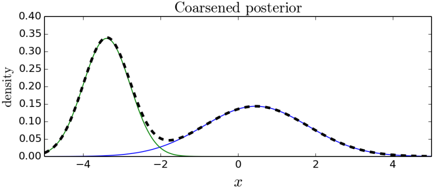
Gaussian mixture applied to skew-normal mixture data


Data distribution

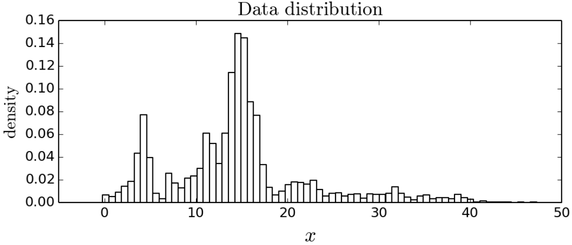

Standard posterior


Standard posterior


Coarsened posterior

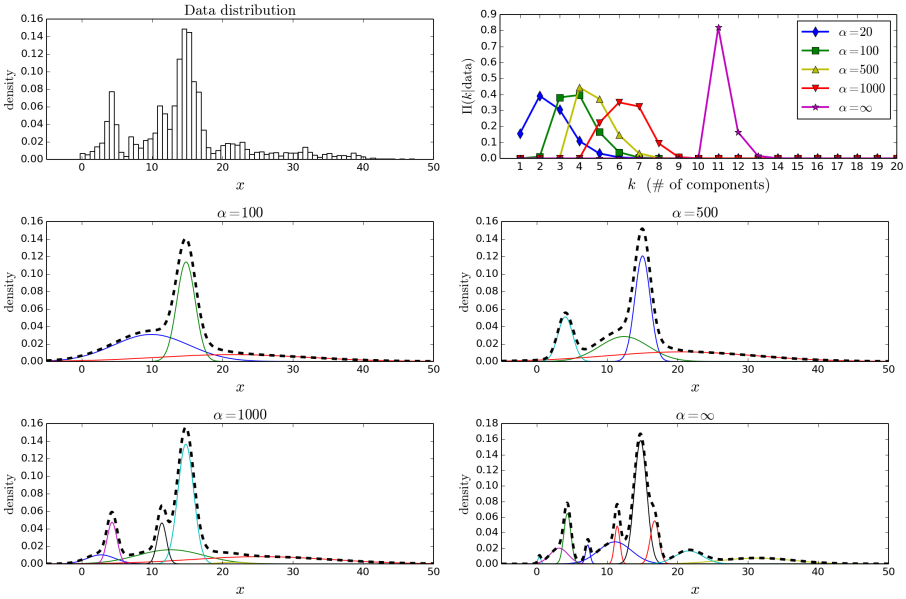

Coarsened posterior

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.
- By considering a range of α values, we can explore the data at varying levels of precision.

Velocities of galaxies in the Shapley supercluster



Thank you!

Inference using Partial Information

Jeff Miller

Harvard University
Department of Biostatistics

ICERM Probabilistic Scientific Computing workshop
June 8, 2017