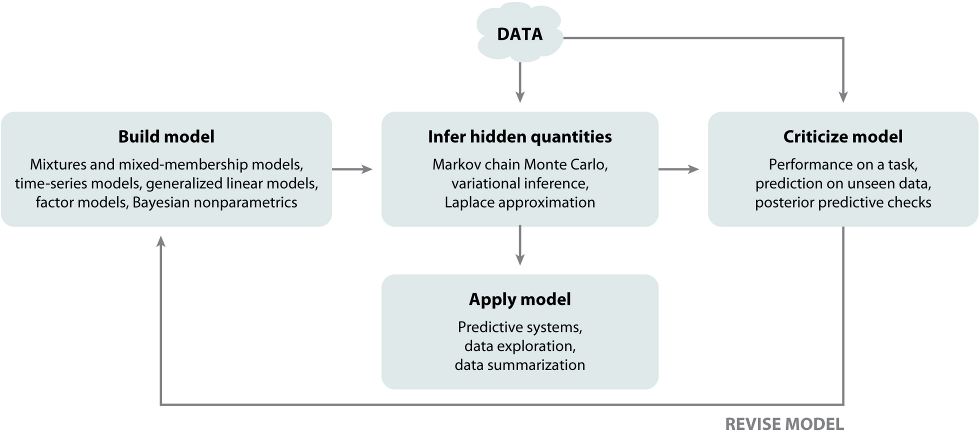
STATS271/371: Applied Bayesian Statistics

Bayesian Mixture Models and (Collapsed) Gibbs Sampling

Scott Linderman

April 26, 2021

Box’s Loop



DATA

Build model

Mixtures and mixed-membership models, time-series models, generalized linear models, factor models, Bayesian nonparametrics

Infer hidden quantities

Markov chain Monte Carlo, variational inference, Laplace approximation

Criticize model

Performance on a task, prediction on unseen data, posterior predictive checks

Apply model

Predictive systems, data exploration, data summarization

REVISE MODEL

Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

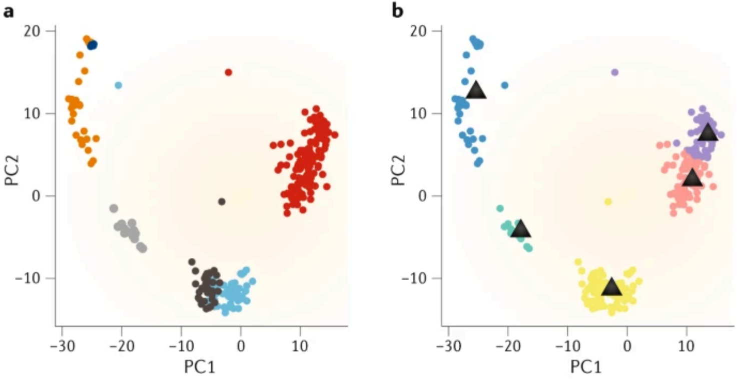
Blei, *Ann. Rev. Stat. App.* 2014.

Lap 4: Bayesian Mixture Models and (Collapsed) Gibbs Sampling

- **Model:** Bayesian mixture models
- **Algorithm:** Gibbs sampling
- **Criticism:** Posterior predictive checks
- **Algorithm II:** Collapsed Gibbs sampling

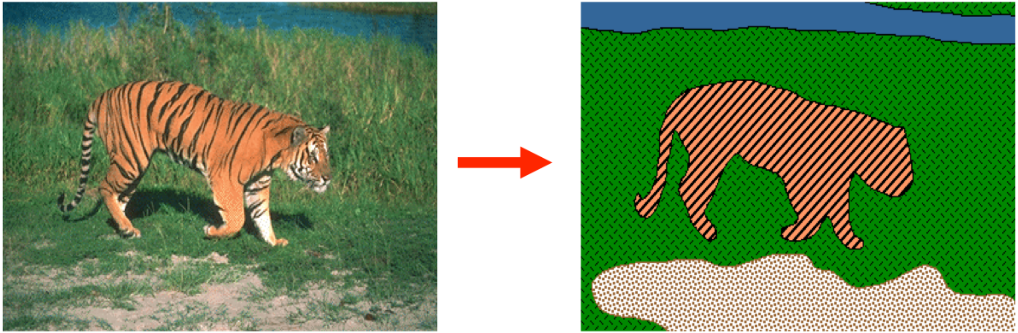
Motivation: Clustering scRNA-seq data

Fig. 3: Clustering methods for scRNA-seq.



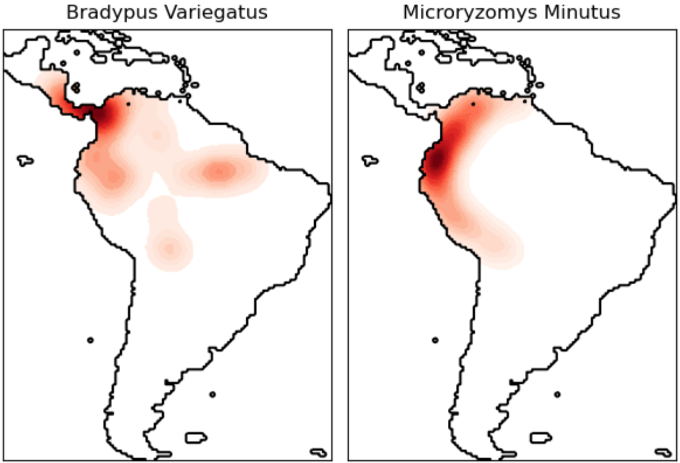
From Kiselev et al. [2019]

Motivation: Foreground/background segmentation



https://ai.stanford.edu/~syyeung/cvweb/tutorial3.html

Motivation: Density estimation



https://scikit-learn.org/stable/auto_examples/neighbors/plot_species_kde.html

Notation

Constants: Let

- N denote the number of data points.
- K denote the number of mixture components (i.e. clusters)

Data: Let

- $\boldsymbol{x}_n \in \mathbb{R}^D$ denote the n-th data point.

Latent Variables: Let

- $z_n \in \{1, \ldots, K\}$ denote the *assignment* of the n-th data point.

Notation II

Parameters: Let

- $\boldsymbol{\eta}_k$ denote the *natural parameters* of component k
- $\boldsymbol{\pi} \in \Delta_K$ denote the component *proportions* (i.e. probabilities).

Hyperparameters: Let

- $\boldsymbol{\phi}, \nu$ denote hyperparameters of the prior on $\boldsymbol{\eta}$
- $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ denote the concentration of the prior on proportions.

Generative Model

1. Sample the proportions from a Dirichlet prior:

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{1}$$

2. Sample the parameters for each component:

$$\boldsymbol{\eta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\eta} \mid \boldsymbol{\phi}, \nu) \qquad \text{for } k = 1, \ldots, K \tag{2}$$

3. Sample the assignment of each data point:

$$z_n \overset{\text{iid}}{\sim} \boldsymbol{\pi} \qquad \text{for } n = 1, \ldots, N \tag{3}$$

4. Sample data points given their assignments:

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\eta}_{z_n}) \qquad \text{for } n = 1, \ldots, N \tag{4}$$

Joint distribution

This generative model corresponds to the following factorization of the joint distribution,

$$p(\pi, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N p(z_n \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \{\boldsymbol{\eta}_k\}_{k=1}^K) \tag{5}$$

Equivalently,

$$p(\pi, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \\ p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K \left[\Pr(z_n = k \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k)\right]^{\mathbb{I}[z_n = k]} \tag{6}$$

Substituting in the assumed forms

$$p(\pi, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K \left[\pi_k\, p(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k)\right]^{\mathbb{I}[z_n = k]} \tag{7}$$

Exponential family mixture models

What about $p(\boldsymbol{x} \mid \boldsymbol{\eta}_k)$ and $p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu)$?

Recall the *exponential family* distributions from Lap 2. Let's assume an exponential family likelihood,

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}_k) = h(\boldsymbol{x}_n) \exp\left\{\langle t(\boldsymbol{x}_n), \boldsymbol{\eta}_k\rangle - A(\boldsymbol{\eta}_k)\right\}. \tag{8}$$

Then assume a *conjugate prior*,

$$p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) \propto \exp\left\{\langle \boldsymbol{\phi}, \boldsymbol{\eta}_k\rangle - \nu A(\boldsymbol{\eta}_k)\right\}. \tag{9}$$

The hyperparmeters $\boldsymbol{\phi}$ are *pseudo-observations* of the sufficient statistics (like statistics from fake data points) and ν is a *pseudo-count* (like the number of fake data points).

Note that the product of prior and likelihood remains in the same family as the prior. That's why we call it conjugate.

Example: Gaussian mixture model

Assume the conditional distribution of $\boldsymbol{x}_n$ is a Gaussian with mean $\boldsymbol{\eta}_{z_n} \in \mathbb{R}^D$ and identity covariance,

$$\begin{aligned} p(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k) &= \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k, \boldsymbol{I}) && (10) \\ &= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\eta}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\eta}_k)\right\} && (11) \\ &= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{x}_n + \boldsymbol{x}_n^\top \boldsymbol{\eta}_k - \tfrac{1}{2}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k\right\}, && (12) \end{aligned}$$

which is an exponential family distribution with base measure $h(\boldsymbol{x}_n) = (2\pi)^{-D/2} e^{-\frac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{x}_n}$, sufficient statistics $t(\boldsymbol{x}_n) = \boldsymbol{x}_n$, and log normalizer $A(\boldsymbol{\eta}_k) = \frac{1}{2}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k$.

Then assume a Gaussian prior on the component parameters. It’s conjugate,

$$p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) = \mathcal{N}(\nu^{-1}\boldsymbol{\phi}, \nu^{-1}\boldsymbol{I}) \propto \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\eta}_k - \tfrac{\nu}{2}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k\right\} = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\eta}_k - \nu A(\boldsymbol{\eta}_k)\right\}. \qquad (13)$$

Note that $\boldsymbol{\phi}$ sets the location and ν sets the precision (i.e. inverse variance).

MAP inference via coordinate ascent

Before diving into fully Bayesian inference algorithms, let's first consider **MAP inference**.

Idea: find the mode of $p(\pi, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{z_n\}_{n=1}^N \mid \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha})$ by **coordinate ascent**.

For now, set $\boldsymbol{\phi} = \mathbf{0}$, and $\nu = 0$ so that the prior is an (improper) uniform distribution. Then maximizing the posterior is equivalent to maximizing the likelihood.

While we're simplifying, let's even fix $\pi = \frac{1}{K}\mathbf{1}_K$.

Coordinate ascent in the Gaussian mixture model

For the Gaussian mixture model (with uniform prior and $\boldsymbol{\pi} = \frac{1}{K}\mathbf{1}_K$), coordinate ascent amounts to:

1. For each $n = 1, \ldots, N$, fix all variables but z_n and find $z_n^\star$ that maximizes

$$p(\boldsymbol{\pi}, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto p(\boldsymbol{x}_n \mid z_n, \{\boldsymbol{\eta}_k\}_{k=1}^K) = \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\eta}_{z_n}, \boldsymbol{I}) \tag{14}$$

   The cluster assignment that maximizes the likelihood is the one with the closest mean to $\boldsymbol{x}_n$, so set

$$z_n^\star = \underset{k \in \{1,\ldots,K\}}{\arg\min} \|\boldsymbol{x}_n - \boldsymbol{\eta}_k\|_2. \tag{15}$$

Coordinate ascent in the Gaussian mixture model II

2 For each $k = 1, \dots, K$, fix all variables but $\boldsymbol{\eta}_k$ and find $\boldsymbol{\eta}_k^\star$ that maximizes,

$$p(\boldsymbol{\pi}, \{\boldsymbol{\eta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \prod_{n=1}^N p(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k)^{\mathbb{I}[z_n=k]} \tag{16}$$

$$\propto \exp\left\{\sum_{n=1}^N \mathbb{I}[z_n = k]\left(\boldsymbol{x}_n^\top \boldsymbol{\eta}_k - \tfrac{1}{2}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k\right)\right\} \tag{17}$$

Taking the derivative of the log and setting to zero yields,

$$\boldsymbol{\eta}_k^\star = \frac{1}{N_k}\sum_{n=1}^K \mathbb{I}[z_n = k]\boldsymbol{x}_n, \tag{18}$$

where $N_k = \sum_{n=1}^N \mathbb{I}[z_n = k]$.

This is the **k-means algorithm**!

Aside: EM in the Gaussian mixture model

We'll talk more about *coordinate ascent variational inference* (CAVI) and *expectation-maximization* (EM) next week. Not to spoil the surprise, but we'll see that they have a similar flavor. Instead of assigning $z_n^\star$ to the closest cluster, we compute *responsibilities* for each cluster:

1. For each data point n and component k, set the *responsibility* to,

$$\omega_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k, \boldsymbol{I})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\eta}_j, \boldsymbol{I})}. \tag{19}$$

2. For each component k, set the mean to

$$\boldsymbol{\eta}_k^\star = \frac{1}{N_k} \sum_{n=1}^{K} \omega_{nk} \boldsymbol{x}_n, \tag{20}$$

   where $N_k = \sum_{n=1}^N \omega_{nk}$.

Note that EM allows for arbitrary proportions $\boldsymbol{\pi}$. Those can be updated as well: for each component k, set $\pi_k = \frac{N_k}{N}$.

Lap 4: Bayesian Mixture Models and (Collapsed) Gibbs Sampling

- Model: Bayesian mixture models
- **Algorithm: Gibbs sampling**
- Criticism: Posterior predictive checks
- Algorithm II: Collapsed Gibbs sampling

Gibbs sampling in Bayesian mixture models

Idea: just like in coordinate ascent, update one variable at a time. *But rather than setting it to its conditional mode, sample from its conditional distribution.*

Gibbs sampling in Bayesian mixture models II

1. For each data point *n*, sample a new assignment from the complete conditional distribution

$$z_n \sim p(z_n \mid \{\boldsymbol{x}_n\}_{n=1}^N, \{z_{n'}\}_{n' \neq n}, \{\boldsymbol{\eta}_k\}_{k=1}^K, \boldsymbol{\pi}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}). \tag{21}$$

Thanks to the factorization of the joint distribution,

$$\Pr(z_n = k \mid -) \propto \Pr(z_n = k \mid \boldsymbol{\pi})\, p(x_n \mid \boldsymbol{\eta}_k) \tag{22}$$

In the Gaussian mixture model, this is,

$$\Pr(z_n = k \mid -) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k, \boldsymbol{I})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\eta}_j, \boldsymbol{I})} \tag{23}$$

$$\equiv \omega_{nk}. \tag{24}$$

I.e., Gibbs sampling generates *random* assignments by sampling according to the responsibilities.

Gibbs sampling in Bayesian mixture models III

2 For each component k, sample new parameters from their complete conditional,

$$\boldsymbol{\eta}_k \sim p(\boldsymbol{\eta}_k \mid \{(\boldsymbol{x}_n, z_n)\}_{n=1}^N, \{\boldsymbol{\eta}_{k'}\}_{k' \neq k}, \boldsymbol{\pi}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}). \tag{25}$$

Thanks to the factorization of the joint distribution,

$$p(\boldsymbol{\eta}_k \mid -) \propto p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n:z_n=k} p(x_n \mid \boldsymbol{\eta}_k). \tag{26}$$

In an Gaussian mixture model,

$$p(\boldsymbol{\eta}_k \mid -) \propto \exp\left\{\left(\boldsymbol{\phi} + \sum_{n=1}^N \mathbb{I}[z_n = k]\boldsymbol{x}_n\right)^\top \boldsymbol{\eta}_k - \frac{\nu + N_k}{2}\boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k\right\} \tag{27}$$

$$\propto \mathcal{N}\left(\boldsymbol{\eta}_k \mid (\nu + N_k)^{-1}\left(\boldsymbol{\phi} + \sum_{n=1}^N \mathbb{I}[z_n = k]\boldsymbol{x}_n\right), (\nu + N_k)^{-1}\boldsymbol{I}\right) \tag{28}$$

where $N_k = \sum_{n=1}^N \mathbb{I}[z_n = k]$. What happens when $N_k \to \infty$?

Gibbs sampling in Bayesian mixture models IV

3 Finally, sample new component proportions from their complete conditional,

$$\boldsymbol{\pi} \sim p(\boldsymbol{\pi} \mid \{(\boldsymbol{x}_n, z_n)\}_{n=1}^N, \{\boldsymbol{\eta}_k\}_{k=1}^K, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}). \tag{29}$$

Thanks to the factorization of the joint distribution,

$$\begin{aligned} p(\boldsymbol{\pi} \mid -) &\propto \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{n=1}^N p(z_n \mid \boldsymbol{\pi}) && (30) \\ &\propto \prod_{k=1}^K \pi_k^{\alpha_k - 1} \times \prod_{n=1}^N \prod_{k=1}^K \pi_k^{\mathbb{I}[z_n = k]} && (31) \\ &\propto \mathrm{Dir}\left(\boldsymbol{\pi} \mid [\alpha_1 + N_1, \ldots, \alpha_K + N_K]\right) && (32) \end{aligned}$$

where $N_k = \sum_{n=1}^N \mathbb{I}[z_n = k]$. What happens when $N_k \to \infty$?

Gibbs sampling in Bayesian exponential family mixture models

What happens in general exponential family models? Step 2 becomes,

2 For each component k, sample new parameters from their complete conditional,

$$p(\boldsymbol{\eta}_k \mid -) \propto \exp\left\{\left\langle \boldsymbol{\phi} + \sum_{n=1}^{N} \mathbb{I}[z_n = k] t(\boldsymbol{x}_n), \boldsymbol{\eta}_k \right\rangle - (\nu + N_k) A(\boldsymbol{\eta}_k)\right\} \tag{33}$$

$$= p\left(\boldsymbol{\eta}_k \,\middle|\, \boldsymbol{\phi} + \sum_{n=1}^{N} \mathbb{I}[z_n = k]\, t(\boldsymbol{x}_n), \nu + N_k\right) \tag{34}$$

where $N_k = \sum_{n=1}^{N} \mathbb{I}[z_n = k]$.

Opportunities for parallelism

- In all three algorithms above, the updates of z_n are independent of one another (once you fix $\{\boldsymbol{\eta}_k\}_{k=1}^K$) and hence can be performed in parallel.
- Likewise, the updates of $\boldsymbol{\eta}_k$ are independent of one another (once you fix $\{z_n\}_{n=1}^N$) and hence can be performed in parallel.
- In fact, we can write these as simple map-reduce algorithms and take advantage of parallel hardware if it's available.
- In the Gibbs sampling case, updating many variables at once from their combined conditional distribution is called **blocked Gibbs sampling**, and it's particularly easy when the variables are conditionally independent, as in the Bayesian mixture model.

Why does Gibbs sampling work?

Gibbs is a special case of MH with proposals that always accept.

Consider a more general setting with parameters $\boldsymbol{\theta} \in \mathbb{R}^P$ and dataset $\mathscr{D}$. In the mixture model, $\boldsymbol{\theta} = (\{z_n\}_{n=1}^N, \{\boldsymbol{\eta}_k\}_{k=1}^K, \boldsymbol{\pi})$ and $\mathscr{D} = \{\boldsymbol{x}_n\}_{n=1}^N$

Gibbs sampling updates one "coordinate" of $\boldsymbol{\theta}$ at a time by sampling from its conditional distribution.

Think of this as a proposal distribution. For each coordinate $j \in 1, \ldots, P$,

$$q_j(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = p(\theta_j \mid \boldsymbol{\theta}'_{\neg j}, \mathscr{D})\, \delta_{\boldsymbol{\theta}'_{\neg j}}(\boldsymbol{\theta}_{\neg j}), \tag{35}$$

where $\boldsymbol{\theta}_{\neg j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_P)$ denotes all parameters except θ_j.

In other words, the proposal distribution q_j samples θ_j from its conditional distribution and leaves all the other parameters unchanged.

Why does Gibbs sampling work? II

What is the probability of accepting this proposal?

$$a_j(\boldsymbol{\theta}' \to \boldsymbol{\theta}) = \min\left\{1,\ \frac{p(\boldsymbol{\theta}, \mathscr{D}) q_j(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta}', \mathscr{D}) q_j(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}\right\} \tag{36}$$

$$= \min\left\{1,\ \frac{p(\boldsymbol{\theta}, \mathscr{D}) p(\theta_j' \mid \boldsymbol{\theta}_{\neg j}, \mathscr{D}) \delta_{\boldsymbol{\theta}_{\neg j}}(\boldsymbol{\theta}'_{\neg j})}{p(\boldsymbol{\theta}', \mathscr{D}) p(\theta_j \mid \boldsymbol{\theta}'_{\neg j}, \mathscr{D}) \delta_{\boldsymbol{\theta}'_{\neg j}}(\boldsymbol{\theta}_{\neg j})}\right\} \tag{37}$$

$$= \min\left\{1,\ \frac{p(\boldsymbol{\theta}_{\neg j}, \mathscr{D}) p(\theta_j \mid \boldsymbol{\theta}_{\neg j}, \mathscr{D}) p(\theta_j' \mid \boldsymbol{\theta}_{\neg j}, \mathscr{D}) \delta_{\boldsymbol{\theta}_{\neg j}}(\boldsymbol{\theta}'_{\neg j})}{p(\boldsymbol{\theta}'_{\neg j}, \mathscr{D}) p(\theta_j' \mid \boldsymbol{\theta}'_{\neg j}, \mathscr{D}) p(\theta_j \mid \boldsymbol{\theta}'_{\neg j}, \mathscr{D}) \delta_{\boldsymbol{\theta}'_{\neg j}}(\boldsymbol{\theta}_{\neg j})}\right\} \tag{38}$$

$$= \min\{1, 1\} = 1 \tag{39}$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$ that differ only in their j-th coordinate.

The Gibbs proposal is *an offer you cannot refuse*.

Why does Gibbs sampling work? III

Of course, if we only update one coordinate, the chain can’t be ergodic. However, if we cycle through coordinates it generally will be.

Question: Does the order in which we update coordinates matter?

Metropolis-Hastings within Gibbs

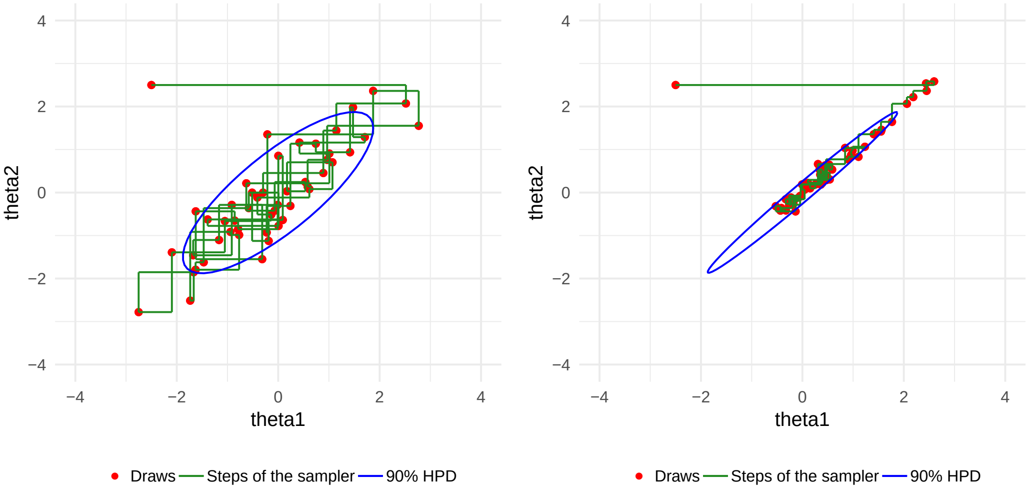
What if we cannot exactly sample the conditional distribution for some coordinates?

We can mix and match Gibbs and MH updates as long as each update preserves the stationary distribution and the collection of transitions forms an ergodic chain.

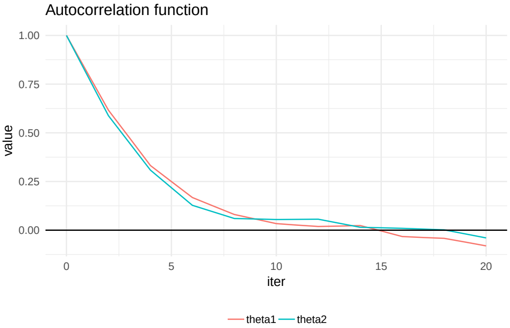
Example: suppose $p(\boldsymbol{\eta}_k \mid \{z_n, \boldsymbol{x}_n\}_{n=1}^N, \{\boldsymbol{\eta}_{k'}\}_{k' \neq k}, \boldsymbol{\pi})$ could not be sampled exactly. If $\boldsymbol{\eta}_k$ were a continuous r.v. with a differentiable log conditional density, we could apply HMC to update $\boldsymbol{\eta}_k$ and Gibbs to update other variables.

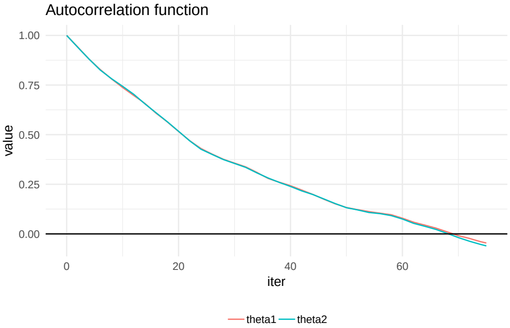
Part of the "art" of applied Bayesian statistics is designing MCMC transitions to effectively sample the posterior distribution, leveraging model structure (exact conditionals, differentiability, etc.) where possible.

Gibbs in a 2D Gaussian example



Gibbs in a 2D Gaussian example


Autocorrelation function


Autocorrelation function

Lap 4: Bayesian Mixture Models and (Collapsed) Gibbs Sampling

- Model: Bayesian mixture models
- Algorithm: Gibbs sampling
- **Criticism: Posterior predictive checks**
- Algorithm II: Collapsed Gibbs sampling

Slides for this section are adapted from Aki Vehtari's lecture notes.

`https://github.com/avehtari/BDA_course_Aalto/blob/master/slides/`

Chapter 6 of BDA3: Model checking

- 6.1 The place of model checking in applied Bayesian statistics
- 6.2 Do the inferences from the model make sense?
- 6.3 Posterior predictive checking
- 6.4 Graphical posterior predictive checks (can be skipped)
- 6.5 Model checking for the educational testing example

Model checking: Overview

- Sensibility with respect to additional information not used in modeling
  - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
  - compare predictions to completely new observations
  - cf. relativity theory predictions
- Internal validation
  - posterior predictive checking
  - cross-validation predictive checking

Model checking: Overview

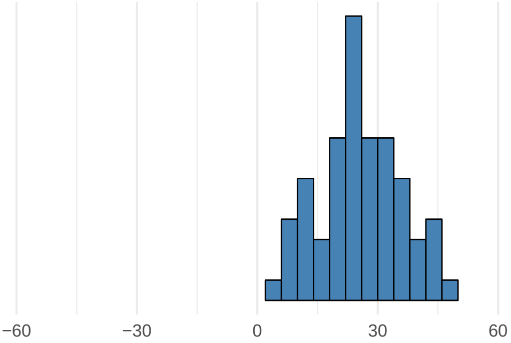
- Sensibility with respect to additional information not used in modeling
    - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
    - compare predictions to completely new observations
    - cf. relativity theory predictions
- Internal validation
    - posterior predictive checking
    - cross-validation predictive checking

Model checking: Overview

- Sensibility with respect to additional information not used in modeling
    - e.g., if posterior would claim that hazardous chemical decreases probability of death
- External validation
    - compare predictions to completely new observations
    - cf. relativity theory predictions
- Internal validation
    - posterior predictive checking
    - cross-validation predictive checking

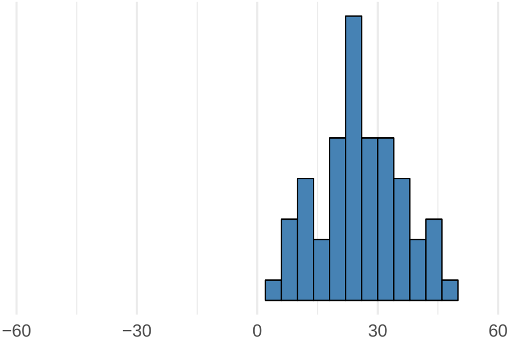
Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log\sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma \mid y)$
  - draw $y^{\mathrm{rep}(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates



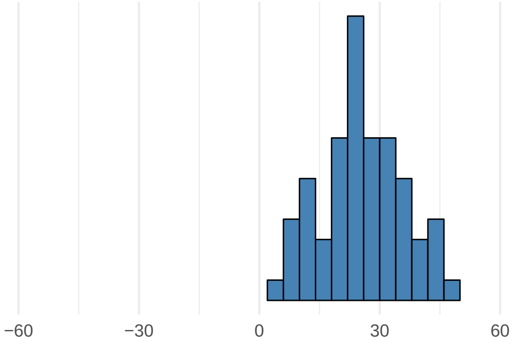
Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma \mid y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates



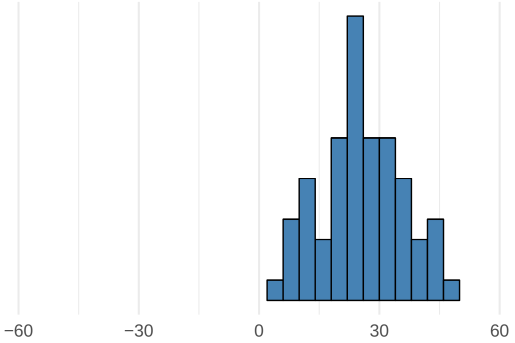
Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma \mid y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates



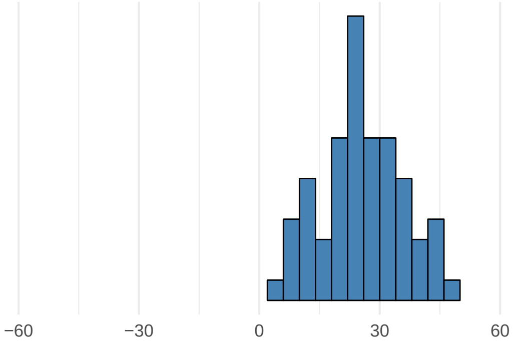
Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma \mid y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates



Posterior predictive checking – example

- Newcomb's speed of light measurements
  - model $y \sim \mathcal{N}(\mu, \sigma)$ with prior $(\mu, \log \sigma) \propto 1$
- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma \mid y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates

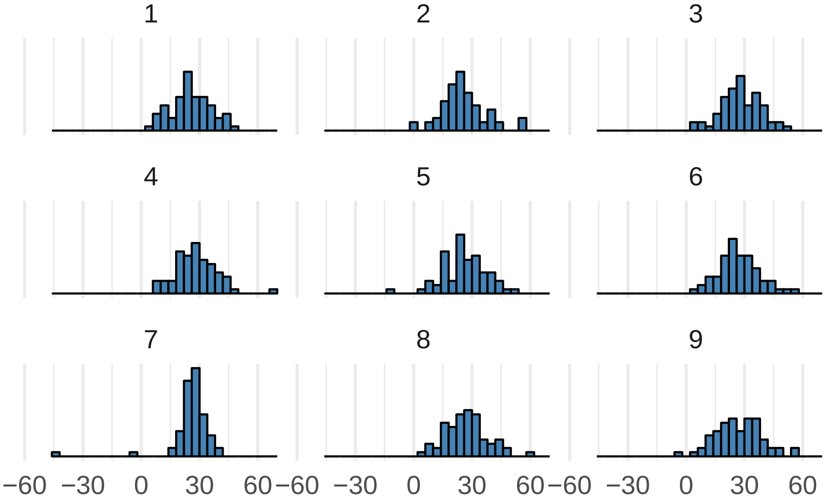


Replicates vs. future observation

- Predictive $\tilde{y}$ is the next not yet observed possible observation. y^{rep} refers to replicating the whole experiment (potentially with same values of x) and obtaining as many replicated observations as in the original data.

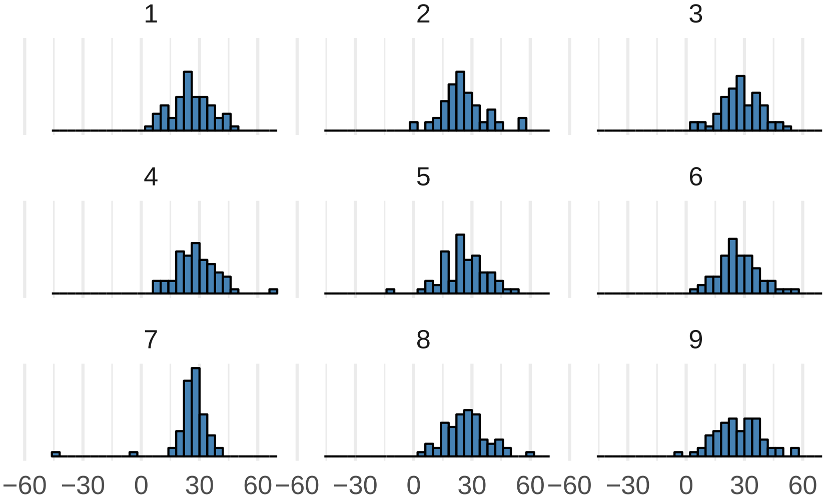
Posterior predictive checking – example

- Generate several replicated datasets y^{rep}
- Compare to the original dataset



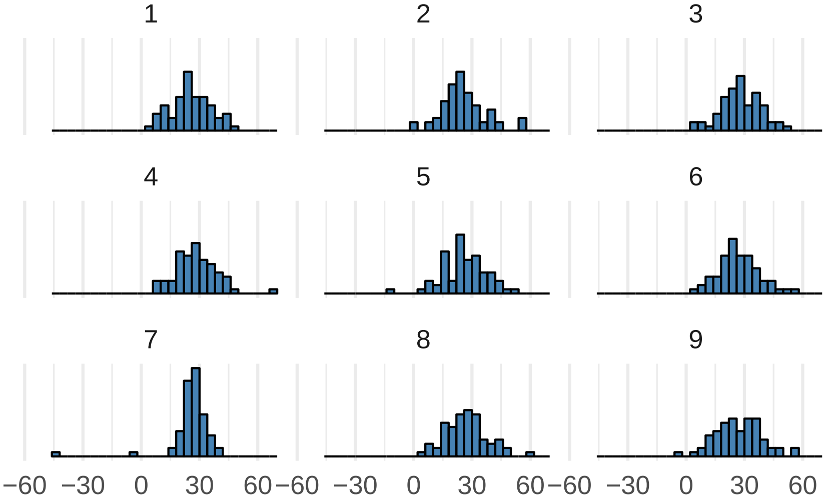
Posterior predictive checking – example

- Generate several replicated datasets y^{rep}
- Compare to the original dataset



Posterior predictive checking – example

- Generate several replicated datasets y^{rep}
- Compare to the original dataset

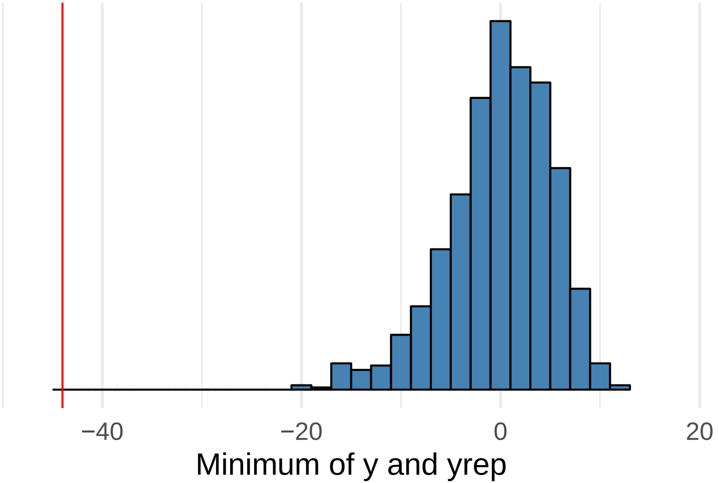


Posterior predictive checking with test statistic

- Replicated data sets y^{rep}
- Test quantity (or discrepancy measure) $T(y, \theta)$
  - summary quantity for the observed data $T(y, \theta)$
  - summary quantity for a replicated data $T(y^{\text{rep}}, \theta)$
  - can be easier to compare summary quantities than data sets

Posterior predictive checking – example

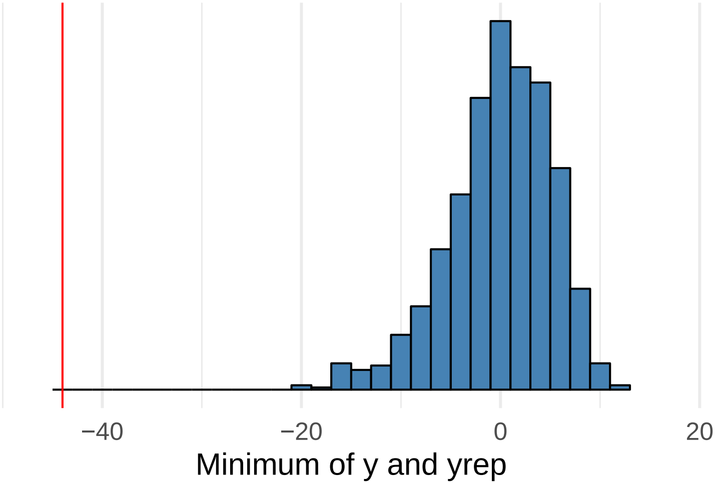
- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\mathrm{rep}})$ for many replicated datasets



Posterior predictive checking – example

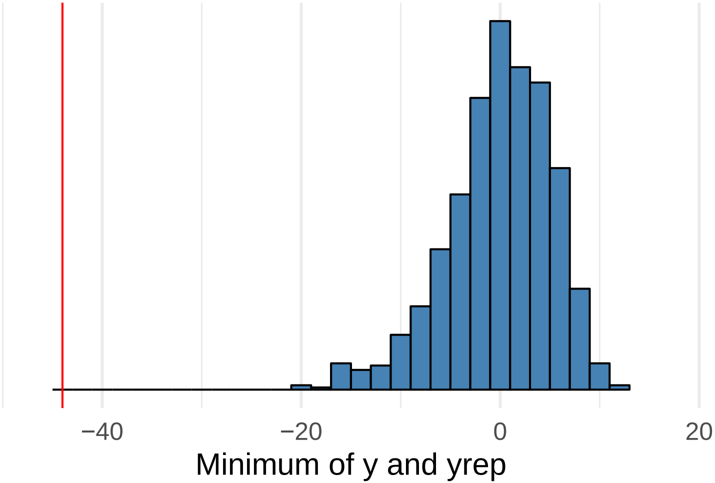
- Compute test statistic for data $T(y, \theta) = \min(y)$

- Compute test statistic $\min(y^{\text{rep}})$ for many replicated datasets



Posterior predictive checking – example

- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\text{rep}})$ for many replicated datasets



Posterior predictive checking

- *Posterior predictive p-value*

$$
\begin{aligned}
p &= \Pr(T(y^{\text{rep}},\theta) \geq T(y,\theta) \mid y) \\
&= \int\int \mathbb{I}[T(y^{\text{rep}},\theta) \geq T(y,\theta)]\, p(y^{\text{rep}} \mid \theta) p(\theta \mid y) dy^{\text{rep}} d\theta
\end{aligned}
$$

  where I is an indicator function

  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$
T(y^{\text{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S
$$

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance

- Not commonly used, since the distribution of test statistic has more information

Posterior predictive checking

- *Posterior predictive p-value*

$$\begin{aligned} p &= \Pr(T(y^{\text{rep}},\theta) \geq T(y,\theta) \mid y) \\ &= \int\int \mathbb{I}[T(y^{\text{rep}},\theta) \geq T(y,\theta)]\, p(y^{\text{rep}} \mid \theta) p(\theta \mid y) dy^{\text{rep}} d\theta \end{aligned}$$

  where I is an indicator function

  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$T(y^{\text{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance

- Not commonly used, since the distribution of test statistic has more information

Posterior predictive checking

- *Posterior predictive p-value*

$$\begin{aligned} p &= \Pr(T(y^{\text{rep}},\theta) \geq T(y,\theta) \mid y) \\ &= \int\int \mathbb{I}[T(y^{\text{rep}},\theta) \geq T(y,\theta)]\, p(y^{\text{rep}} \mid \theta) p(\theta \mid y) dy^{\text{rep}} d\theta \end{aligned}$$

  where I is an indicator function

  - having $(y^{\text{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$T(y^{\text{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance
- Not commonly used, since the distribution of test statistic has more information

Posterior predictive checking

- *Posterior predictive p-value*

$$\begin{aligned} p &= \Pr(T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta) \mid y) \\ &= \int\int \mathbb{I}[T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta)] p(y^{\mathrm{rep}} \mid \theta) p(\theta \mid y) dy^{\mathrm{rep}} d\theta \end{aligned}$$

  where I is an indicator function

  - having $(y^{\mathrm{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

$$T(y^{\mathrm{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance

- Not commonly used, since the distribution of test statistic has more information

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. *t* distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to “outliers”
    - e.g. *t* instead of Gaussian
- Compare sensitivity of essential inference quantities
  - extreme quantiles are more sensitive than means and medians
  - extrapolation is more sensitive than interpolation

Sensitivity analysis

- How much different choices in model structure and priors affect the results
    - test different models and priors
    - alternatively combine different models to one model
        - e.g. hierarchical model instead of separate and pooled
        - e.g. *t* distribution contains Gaussian as a special case
    - robust models are good for testing sensitivity to “outliers”
        - e.g. *t* instead of Gaussian
- Compare sensitivity of essential inference quantities
    - extreme quantiles are more sensitive than means and medians
    - extrapolation is more sensitive than interpolation

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. *t* distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to “outliers”
    - e.g. *t* instead of Gaussian
- Compare sensitivity of essential inference quantities
  - extreme quantiles are more sensitive than means and medians
  - extrapolation is more sensitive than interpolation

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. *t* distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to "outliers"
    - e.g. *t* instead of Gaussian
- Compare sensitivity of essential inference quantities
  - extreme quantiles are more sensitive than means and medians
  - extrapolation is more sensitive than interpolation

Lap 4: Bayesian Mixture Models and (Collapsed) Gibbs Sampling

- Model: Bayesian mixture models
- Algorithm: Gibbs sampling
- Criticism: Posterior predictive checks
- **Algorithm II: Collapsed Gibbs sampling**

"Collapsing" out variables

In some models, we can marginalize (aka *collapse* or *integrate out*) some variables to work on a lower dimensional distribution.

Typically, this is possible in models constructed with conjugate exponential family distributions.

Collapsed Gibbs for Bayesian mixtures

Let's marginalize the parameters $\{\boldsymbol{\eta}_k\}_{k=1}^K$ in the exponential family mixture model,

$$p(\pi, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K \left[\int p(\boldsymbol{\eta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N [\pi_k \, p(\boldsymbol{x}_n \mid \boldsymbol{\eta}_k)]^{\mathbb{I}[z_n = k]} \, \mathrm{d}\boldsymbol{\eta}_k \right] \tag{40}$$

$$\propto \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K \left[\pi_k^{N_k} \int \frac{1}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}, \nu)} \exp \left\{ \left\langle \boldsymbol{\phi} + \sum_{n: z_n = k} t(\boldsymbol{x}_n), \boldsymbol{\eta}_k \right\rangle - (\nu + N_k) A(\boldsymbol{\eta}_k) \right\} \mathrm{d}\boldsymbol{\eta}_k \right] \tag{41}$$

$$= \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K \left[\pi_k^{N_k} \frac{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi} + \sum_{n: z_n = k} t(\boldsymbol{x}_n), \nu + N_k)}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}, \nu)} \right] \tag{42}$$

where $Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}, \nu)$ is the normalizing function of the conjugate prior $p(\boldsymbol{\eta} \mid \boldsymbol{\phi}, \nu)$.

Question: can we still parallelize the Gibbs updates of $\{z_n\}_{n=1}^N$?

Collapsed Gibbs for Bayesian mixtures II

While we're at it, let's marginalize the mixture proportions $\boldsymbol{\pi}$, too. The Dirichlet density is,

$$\mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{1}{Z_\pi(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \quad \text{where} \quad Z_\pi(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{43}$$

Plugging this in and integrating over $\boldsymbol{\pi}$ yields,

$$p(\{(z_n, \boldsymbol{x}_n)\}_{n=1}^{N} \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \left[\int \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^{K} \pi_k^{N_k} \, \mathrm{d}\boldsymbol{\pi}\right] \left[\prod_{k=1}^{K} \frac{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n), \nu + N_k)}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}, \nu)}\right] \tag{44}$$

$$= \left[\frac{Z_\pi([\alpha_1 + N_1, \ldots, \alpha_K + N_K])}{Z_\pi(\boldsymbol{\alpha})}\right] \left[\prod_{k=1}^{K} \frac{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n), \nu + N_k)}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}, \nu)}\right] \tag{45}$$

Collapsed Gibbs for Bayesian Mixtures III

We'll simplify the notation by writing,

$$p(\{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha})} \prod_{k=1}^K \frac{Z_\eta(\boldsymbol{\phi}'_k, \nu'_k)}{Z_\eta(\boldsymbol{\phi}, \nu)} \tag{46}$$

where

$$\boldsymbol{\alpha}' = [\alpha_1 + N_1, \ldots, \alpha_K + N_K] \tag{47}$$

$$\boldsymbol{\phi}'_k = \boldsymbol{\phi} + \sum_{n:z_n=k} t(\boldsymbol{x}_n) \tag{48}$$

$$\nu'_k = \nu + N_k. \tag{49}$$

This is a **general pattern**: in exponential families, marginal likelihoods are given by ratios of posterior and prior normalizing functions.

Collapsed Gibbs for Bayesian Mixtures IV

Now consider the conditional distribution of z_n, holding all the other assignments fixed,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto Z_\pi(\boldsymbol{\alpha}') \prod_{k=1}^{K} Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}'_k, \nu'_k) \tag{50}$$

where $\boldsymbol{\alpha}'$, $\boldsymbol{\phi}'_k$, and ν'_k are computed with $z_n = k$. To simplify, divide by a constant w.r.t. z_n,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha}'^{(\neg n)})} \prod_{k=1}^{K} \frac{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}'_k, \nu'_k)}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}_k'^{(\neg n)}, \nu_k'^{(\neg n)})} \tag{51}$$

where

$$\boldsymbol{\alpha}'^{(\neg n)} = [\alpha_1 + N_1^{(\neg n)}, \ldots, \alpha_K + N_K^{(\neg n)}] \qquad \boldsymbol{\phi}_k'^{(\neg n)} = \boldsymbol{\phi} + \sum_{n' \neq n} t(\boldsymbol{x}_{n'}) \mathbb{I}[z_{n'} = k] \tag{52}$$

$$\nu_k'^{(\neg n)} = \nu + N_k^{(\neg n)} \qquad N_k^{(\neg n)} = \sum_{n' \neq n} \mathbb{I}[z_{n'} = k] \tag{53}$$

Collapsed Gibbs for Bayesian Mixtures V

Then many terms cancel. In the first ratio,

$$\frac{Z_\pi(\boldsymbol{\alpha}')}{Z_\pi(\boldsymbol{\alpha}'^{(\neg n)})} = \frac{\prod_{k=1}^K \Gamma(\alpha'_k)\,\Gamma(\sum_{k=1}^K \alpha_k'^{(\neg n)})}{\prod_{k=1}^K \Gamma(\alpha_k'^{(\neg n)})\,\Gamma(\sum_{k=1}^K \alpha'_k)} \propto \alpha_k'^{(\neg n)} = \alpha + N_k^{(\neg n)} \tag{54}$$

In words, the first ratio is proportion to the size of cluster *k* before adding the *n*-th data point.

Consider the second ratio. All but the *k*-th term in the product cancel to leave:

$$\prod_{k=1}^K \frac{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}'_k, \nu'_k)}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}_k'^{(\neg n)}, \nu_k'^{(\neg n)})} = \frac{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}'_k, \nu'_k)}{Z_{\boldsymbol{\eta}}(\boldsymbol{\phi}_k'^{(\neg n)}, \nu_k'^{(\neg n)})} \propto p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu). \tag{55}$$

In words, the second ratio is proportional to the *posterior predictive density*.

Altogether, the conditional distribution of z_n is,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto (\alpha_k + N_k^{(\neg n)})\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu), \tag{56}$$

a function of the size of the cluster and the likelihood of $\boldsymbol{x}_n$ given other points in that cluster.

The infinite limit: Dirichlet process mixture models

Now consider a special case where $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}_K$ and, loosely speaking, take $K \to \infty$. In this limit, we obtain a **Dirichlet process mixture model**.

There's lots of theory about these Bayesian nonparametric models that we won't touch on today [see Orbanz, 2014]. Instead, just note how the collapsed Gibbs sampling algorithm changes.

The probability of assigning the n-th data point to a non-empty cluster is still,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto (\frac{\alpha}{K} + N_k^{(\neg n)})\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu). \tag{57}$$

But now there are only $K_{\text{used}} = \#\mathrm{unique}(\{z_{n'}\}_{n' \neq n})$ non-empty clusters, and the remaining $K - K_{\text{used}}$ unoccupied clusters each have probability,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \tfrac{\alpha}{K}\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu). \tag{58}$$

The infinite limit: Dirichlet process mixture models II

Since all the empty clusters are equivalent, we can combine them to get,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \\ \propto \begin{cases} (\frac{\alpha}{K} + N_k^{(\neg n)})\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{if } k \in \{1, \ldots, K_{\text{used}}\} \\ (K - K_{\text{used}}) \frac{\alpha}{K}\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k = K_{\text{used}} + 1, \end{cases} \tag{59}$$

where we assume that the cluster labels are permuted after each iteration so that only $k = 1, \ldots, K_{\text{used}}$ are non-empty.

As $K \to \infty$, these updates simplify to the classic collapsed Gibbs updates for DPMMs,

$$p(z_n = k \mid \boldsymbol{x}_n, \{(z_n, \boldsymbol{x}_n)\}_{n' \neq n}, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \\ \propto \begin{cases} N_k^{(\neg n)}\, p(\boldsymbol{x}_n \mid \{\boldsymbol{x}_{n'} : z_{n'} = k\}, \boldsymbol{\phi}, \nu) & \text{if } k \in \{1, \ldots, K_{\text{used}}\} \\ \alpha\, p(\boldsymbol{x}_n \mid \boldsymbol{\phi}, \nu) & \text{if } k = K_{\text{used}} + 1. \end{cases} \tag{60}$$

The infinite limit: Dirichlet process mixture models III

As the Gibbs sampler runs, it has some probability of deleting a cluster (by removing its last data point) and some probability (determined by α) of creating a new cluster with one data point. In this sense, the model is **nonparametric**: it doesn't require you to specify K in advance.

These probabilities are *size-biased*, you're more likely to add a data point to a large cluster.

There are many other ways to arrive at the DPMM:

1. via an stochastic process on partitions called the **Chinese restaurant process (CRP)**
2. as a **random measure** on η with a countably infinite number of weighted atoms, only a finite number of which are used.
3. via a **stick-breaking construction** to get the weights of the random measure.

Orbanz [2014] offers an accessible, book-length treatment of these important models.

References I

Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, 20(5):273–282, May 2019.

Peter Orbanz. Lecture notes on Bayesian nonparametrics. May 2014. URL `http://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf`.