Finite mixture models

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

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

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Introduction

- Datasets are often heterogeneous, in the sense that datapoints tend to fall into groups.
- If the group labels are observed, then they can easily be handled — for instance, by treating them as covariates in regression.
- Meanwhile, if the group labels are unobserved, then we can treat them as latent variables and infer them.
- Introducing latent variables leads to mixture distributions.

Introduction: Terminology

- A *latent variable* is an unobserved random variable in the model.

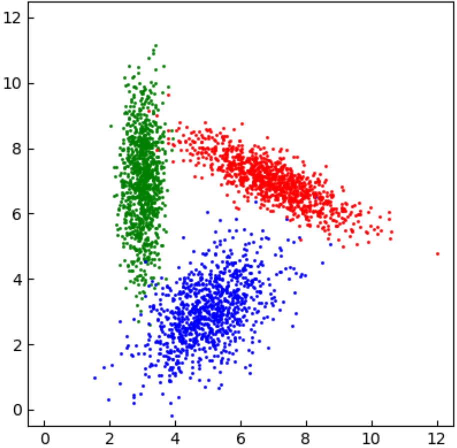
- From the frequentist perspective, latent variables are random and parameters are fixed.

- From the Bayesian perspective, "latent variable" and "parameter" mean essentially the same thing, except:
  - "Parameter" is sometimes used to refer only to continuous latent variables, but this is not a hard-and-fast rule.
  - Latent variables can be discrete or continuous.

- The term "mixture model" usually refers to a mixture in which each datapoint has a discrete latent variable that governs the parameters of the distribution.

Introduction

- Gaussian mixture models are a popular choice, due to their flexibility and computational tractability.



Introduction

- Mixture models can be used for various purposes:
  - Clustering
  - Density estimation
  - Priors on distributions
  - Flexible structured models

- Mixture models are used in many applications:
  - Gene expression profiling (Yeung et al., 2001)
  - Population structure (Pritchard et al., 2000)
  - Computer vision (Stauffer and Grimson, 1999)
  - Speaker recognition (Reynolds et al., 2000)
  - Phylogenetics (Pagel and Meade, 2004)
  - Flow cytometry (Lee and McLachlan, 2014)

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Clustering

Clustering can be used for a wide variety of tasks:

- Finding hidden structure

  e.g., discovering new cancer subtypes

- Summarizing complex data

  e.g., grouping documents on related topics

- Feature construction for supervised learning

  e.g., distance to cluster centers

- Removing unwanted variation

  e.g., population structure in genotype data

- Imputing group labels

  e.g., gating cell types in flow cytometry

K-means clustering

- K-means is a clustering algorithm that is closely related to Gaussian mixture models.

- K-means is one of the oldest and most commonly used clustering algorithms.
- K-means is fast and often works pretty well.

- Basic idea: Initialize by randomly dividing into K groups. Then repeat the following steps until convergence:
  1. Set $\mu_k =$ sample mean of points in group k,
  2. Reassign each point to the group k with the nearest μ_k.

K-means clustering: Demo

(Demo in R)

K-means clustering: Algorithm

K-means algorithm

- Input: Data $x_1, \ldots, x_n \in \mathbb{R}^d$, and an integer $K > 0$.
- Output: Cluster assignments $z_1, \ldots, z_n \in \{1, \ldots, K\}$.

- Randomly initialize $z_1, \ldots, z_n \in \{1, \ldots, K\}$.
- Repeat until no change in the z's is observed:
  1. For $k = 1, \ldots, K$: define $A_k = \{i : z_i = k\}$ and compute
  $$\mu_k \leftarrow \frac{1}{|A_k|} \sum_{i \in A_k} x_i.$$
  2. For $i = 1, \ldots, n$: update $z_i \leftarrow \operatorname{argmin}_k \|x_i - \mu_k\|$.

K-means clustering: Pros and Cons

Pros

- Simple and easy.
- Scales up to large d and large n.
- Converges quickly (i.e., requires few iterations).

Cons

- Sometimes converges to a suboptimal local mode.
- Only makes sense for quantitative data points in $\mathbb{R}^d$.
- Implicitly assumes clusters are radially symmetric and have similar variance.
- We have to choose the number of clusters, K.

Various generalizations can be used to address these disadvantages.

K-means clustering: Mixture model interpretation

- Consider the following model:

$$X_i \sim \mathcal{N}(\mu_{z_i}, I)$$

  indep. for $i = 1, \ldots, n$, where $\mu_k \in \mathbb{R}^d$ and $z_i \in \{1, \ldots, K\}$.

- One way to interpret K-means is that it seeks maximum likelihood estimates of $\mu = (\mu_1, \ldots, \mu_K)$ and $z = (z_1, \ldots, z_n)$.

- Finding the global MLE of μ and z is hard. The likelihood is a complicated function with many local maxima.

- But it is easy to maximize over μ, holding z fixed — just set μ_k equal to the sample average of the x_i's such that $z_i = k$.

- Likewise, it is easy to maximize over z, holding μ fixed — just set $z_i = k$ where μ_k is the nearest μ to x_i.

K-means clustering: Mixture model interpretation

- Thus, K-means alternates between these two maximizations:
  1. Maximize the likelihood over μ, holding z fixed.
  2. Maximize the likelihood over z, holding μ fixed.

- This kind of optimization algorithm is sometimes called "coordinate ascent".
  - Optimize one variable at a time, holding the others fixed.
  - This is analogous to Gibbs sampling, in which we sample each variable given the others, rather than maximizing.

- K-means is guaranteed to increase the likelihood at each iteration — or more precisely, the likelihood never *decreases*.
  - This is true for any coordinate ascent algorithm.

- However, K-means can get stuck in a local maximum.
  - This is usually dealt with by re-running the algorithm many times with different random initializations.

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Gaussian mixture model

- Let's make the z's latent variables by placing priors on them:

$$Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \text{Categorical}(\pi),$$

  that is, $\mathbb{P}(Z_i = k) = \pi_k$, where $\pi_1, \ldots, \pi_K \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$.

- Now, let's generalize to allow component-specific covariances:

$$X_i | z \sim \mathcal{N}(\mu_{z_i}, C_{z_i})$$

  independently for $i = 1, \ldots, n$, where $\mu_k \in \mathbb{R}^d$ and $C_k \in \mathbb{R}^{d \times d}$ is symmetric positive definite for $k = 1, \ldots, K$.

- Equivalently, by marginalizing out the z's,

$$X_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, C_k)$$

  indep. for $i = 1, \ldots, n$. This is a *Gaussian mixture model*.

Gaussian mixtures: Maximum likelihood

- Maximum likelihood estimation of π, μ, and C is hard.
- Expectation–maximization is the most common approach.
- EM for mixtures is similar to K-means, but with weights w rather than binary assignments.

- EM for Gaussian mixtures: Randomly initialize π, μ, and C, then iteratively repeat the following steps:
  1. E-step:
     - $w_{ik} \leftarrow \dfrac{\pi_k \mathcal{N}(x_i \mid \mu_k, C_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, C_j)}$
     - $n_k \leftarrow \displaystyle\sum_{i=1}^{n} w_{ik}$
  2. M-step:
     - $\pi_k \leftarrow n_k / n$
     - $\mu_k \leftarrow \dfrac{1}{n_k} \displaystyle\sum_{i=1}^{n} w_{ik} x_i$
     - $C_k \leftarrow \dfrac{1}{n_k} \displaystyle\sum_{i=1}^{n} w_{ik} (x_i - \mu_k)(x_i - \mu_k)^{\mathtt{T}}$

Gaussian mixtures: Issues with maximum likelihood

- Issue 1: The likelihood has lots of local maxima, and EM tends to get stuck.

- Issue 2: Often, the MLE doesn't even exist.

- The reason is that the likelihood goes to infinity if we set, say, $\mu_1 = x_1$ and take $C_1 \to 0$.

- Issue 2 can be mitigated by putting a lower bound on the scale of each component, but this is kind of hacky.

- Both issues are resolved by using a Bayesian mixture model and MCMC.

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Bayesian Gaussian mixture model (GMM)

- Consider the following model:

$$\begin{aligned} Z_i|\pi,\mu,C &\sim \text{Categorical}(\pi) \\ X_i|z,\pi,\mu,C &\sim \mathcal{N}(\mu_{z_i},C_{z_i}) \end{aligned}$$

  independently for $i = 1, \ldots, n$.

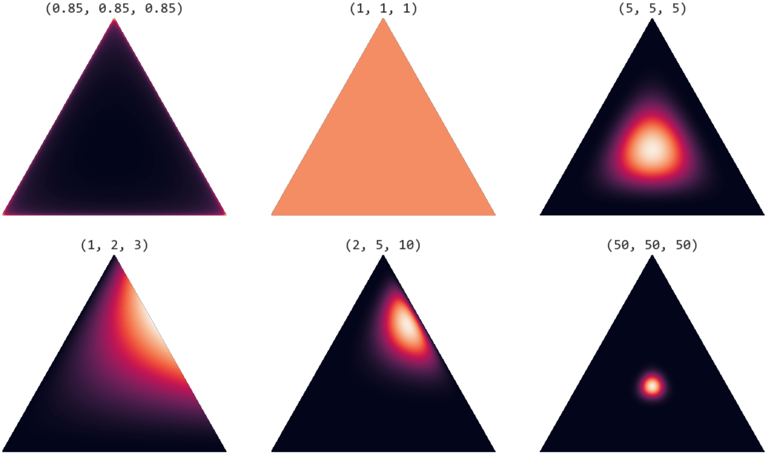
- For brevity, we write $\pi = (\pi_1, \ldots, \pi_K)$, $\mu = (\mu_1, \ldots, \mu_K)$, and $C = (C_1, \ldots, C_K)$.

- We will assume the following priors, independently:

$$\begin{aligned} (\pi_1,\ldots,\pi_K) &\sim \text{Dirichlet}(\alpha_1,\ldots,\alpha_K) \\ \mu_1,\ldots,\mu_K &\overset{\text{iid}}{\sim} \mathcal{N}(m_0,\Sigma_0) \\ C_1,\ldots,C_K &\overset{\text{iid}}{\sim} \text{InverseWishart}(S_0,\nu_0). \end{aligned}$$

- It turns out that these are all semi-conjugate priors.

Aside: Dirichlet distribution



(Image credit: Sue Liu, "Dirichlet distribution: Motivating LDA")

(`https://towardsdatascience.com/dirichlet-distribution-a82ab942a879`)

Aside: Dirichlet distribution

- The Dirichlet distribution is a conjugate prior on the probability vector π in a $\mathrm{Categorical}(\pi)$ distribution.

- It can be thought of as a multivariate version of the Beta distribution, since if $\pi \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$, then $\pi_k \sim \mathrm{Beta}(\alpha_k, \sum_{j \neq k} \alpha_j)$.

- Given $\alpha_1, \ldots, \alpha_K > 0$, the Dirichlet p.d.f. is

$$\mathrm{Dirichlet}(\pi \mid \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

for probability vectors $\pi = (\pi_1, \ldots, \pi_K)$, where

$$B(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}{\Gamma(\alpha_1 + \cdots + \alpha_K)}.$$

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Bayesian GMM: Gibbs sampler (1/3)

- Full conditional for z_i:

$$p(z_i|\cdots) \underset{z_i}{\propto} \text{(Whiteboard activity)}$$

Bayesian GMM: Gibbs sampler (1/3)

- Full conditional for z_i:

$$\begin{aligned} p(z_i|\cdots) &\underset{z_i}{\propto} p(x, z, \pi, \mu, C) \\ &\underset{z_i}{\propto} p(x_i|z, \pi, \mu, C) p(z_i|\pi, \mu, C) \\ &= \mathcal{N}(x_i|\mu_{z_i}, C_{z_i}) \pi_{z_i} \\ &\underset{z_i}{\propto} \mathrm{Categorical}(z_i|w_i) \end{aligned}$$

where $w_i = (w_{i1}, \ldots, w_{iK})$ and $w_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, C_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, C_j)}$.

- Note that the weights w_{ik} are identical to the EM weights.

Bayesian GMM: Gibbs sampler (2/3)

- Full conditional for π:

$$p(\pi | \cdots) \underset{\pi}{\propto} \text{(Whiteboard activity)}$$

Bayesian GMM: Gibbs sampler (2/3)

- Full conditional for π:

$$
\begin{aligned}
p(\pi | \cdots) &\underset{\pi}{\propto} p(x, z, \pi, \mu, C) \\
&\underset{\pi}{\propto} p(z | \pi, \mu, C) p(\pi) \\
&\underset{\pi}{\propto} \Big(\prod_{i=1}^{n} \pi_{z_i} \Big) \Big(\prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \Big) \\
&= \prod_{k=1}^{K} \pi_k^{n_k + \alpha_k - 1} \\
&\underset{\pi}{\propto} \mathrm{Dirichlet}(\pi \mid \alpha_1 + n_1, \ldots, \alpha_K + n_K)
\end{aligned}
$$

where $n_k = \sum_{i=1}^{n} \mathrm{I}(z_i = k)$.

- Here, n_k is defined differently than in EM.

Bayesian GMM: Gibbs sampler (3/3)

- Full conditional for μ_k:

$$
\begin{aligned}
p(\mu_k | \cdots) &\underset{\mu_k}{\propto} p(\mu_k) \prod_{i\,:\,z_i=k} p(x_i | z, \pi, \mu, C) \\
&\underset{\mu_k}{\propto} \mathcal{N}(\mu_k \mid m_0, \Sigma_0) \prod_{i\,:\,z_i=k} \mathcal{N}(x_i \mid \mu_k, C_k) \\
&\underset{\mu_k}{\propto} \mathcal{N}(\mu_k \mid m, \Sigma)
\end{aligned}
$$

where $\Sigma^{-1} = \Sigma_0^{-1} + n_k C_k^{-1}$ and

$$
m = \Sigma\big(\Sigma_0^{-1} m_0 + C_k^{-1} \textstyle\sum_{i\,:\,z_i=k} x_i\big).
$$

- Full conditional for C_k:

$$
p(C_k | \cdots) = \mathrm{InverseWishart}(C_k \mid S, \nu)
$$

where $\nu = \nu_0 + n_k$ and $S = S_0 + \sum_{i\,:\,z_i=k} (x_i - \mu_k)(x_i - \mu_k)^{\mathtt{T}}$.

Using Stan with mixture models

- Stan can only work with continuous parameters, not discrete.
- Thus, Stan cannot sample the latent variables $z_1, \ldots, z_n$.
- Stan's designers recommend using the likelihood with the z's summed out:

$$p(x_1, \ldots, x_n | \pi, \mu, C) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i | \mu_k, C_k).$$

- I've never tried this in Stan, but I'm skeptical of this approach based on my own experience trying this.

Bayesian mixtures of other distributions

- Instead of Gaussians, we can plug in other distributions for the mixture components.
    - Exponential families with conjugate priors are computationally convenient for Gibbs sampling.

- On high-dimensional data, it is useful to constrain the covariance matrices C_k since they are hard to estimate.
    - For example, $C_k = \sigma_k^2 I$ or $C_k = \mathrm{diag}(\sigma_{k1}^2, \ldots, \sigma_{kd}^2)$.
    - The Inverse-Gamma is a conjugate prior on the σ^2's.

- Meanwhile, if a bit more flexibility than Gaussians is desired, the multivariate skew-normal distributions are sometimes useful.

Outline

Introduction

K-means clustering

Gaussian mixture model

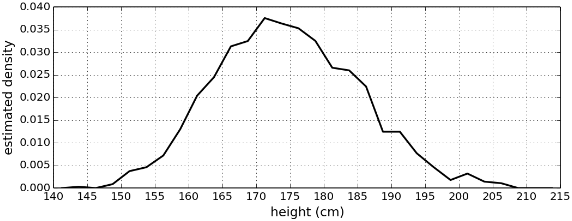
Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Height example



- Let's revisit the example involving the heights of 695 Dutch women and 562 Dutch men.

- Suppose we have the list of heights, but we don't know which datapoints are from women and which are from men.

- Can we still infer the parameters of the female and male distributions separately, e.g., the mean height for each sex?

Height example: Model

- Perhaps surprisingly, the answer is yes.
  - For a finite mixture of Gaussians, it turns out that the parameters are identifiable up to permutation of components.

- In this example, the component assignment variable Z_i indicates whether individual i is female or male.

- For now, to keep things as simple as possible,
  1. assume both components have the same precision, λ, and
  2. assume λ is fixed and known.

- The two-component Gaussian mixture model is

$$\mu_0, \mu_1 \overset{\text{iid}}{\sim} \mathcal{N}(m, \ell^{-1})$$

$$\pi \sim \text{Beta}(a, b)$$

$$Z_1, \ldots, Z_n | \mu, \pi \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi)$$

$X_i | z, \mu, \pi \sim \mathcal{N}(\mu_{z_i}, \lambda^{-1})$ independently for $i = 1, \ldots, n$.

Height example: Hyperparameter settings

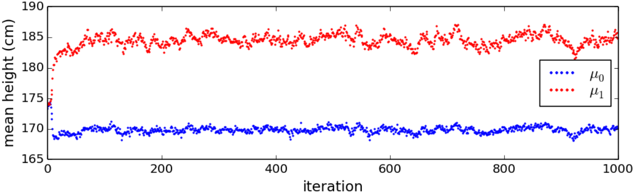
- Let's use the following settings:
  - $\lambda = 1/\sigma^2$ where $\sigma = 8$ cm (≈ 3.1 inches)
    ($\sigma =$ stddev of the subject heights within each component)
  - $a = 1$, $b = 1$ for Beta prior parameters
    (equivalent to prior "sample size" of 1 for each component)
  - $m = 175$ cm (≈ 5' 9")
    (mean of the prior on the component means)
  - $\ell = 1/s^2$ where $s = 15$ cm (≈ 6 inches)
    ($s =$ stddev of the prior on the component means)

Height example: Gibbs sampler settings

- Let's initialize the sampler at:
  - $\pi = 1/2$
    (equal probability for each component)
  - $z_1, \ldots, z_n$ sampled i.i.d. from $\mathrm{Bernoulli}(1/2)$
    (initial assignment to components chosen uniformly at random)
  - $\mu_0 = \mu_1 = m$
    (component means initialized to the mean of their prior)
- Let's do a short run of $N = 10^3$ iterations just for illustration.
  - It probably needs to be run for longer to mix properly.

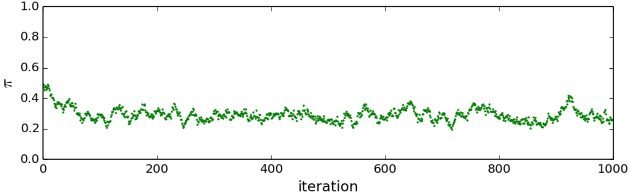
Height example: Results from one Gibbs sampler run

Traceplots of the component means μ_0 and μ_1



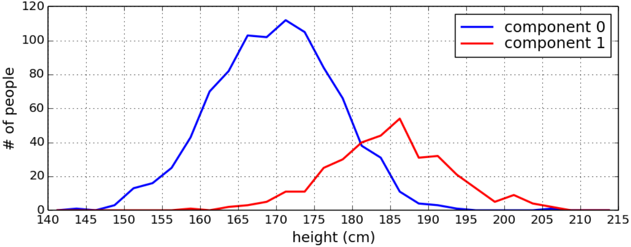
Traceplot of the mixture weight π

($\pi =$ prior probability that a subject comes from component 1)



Height example: Results from one Gibbs sampler run

Histograms of the heights of subjects assigned to each component according to $z_1, \ldots, z_n$ in a typical sample



Height example: Results from one Gibbs sampler run

- From the traceplots of μ_0 and μ_1, we see that one component quickly settles to have a mean of around 168–170 cm and the other to a mean of around 182–186 cm.

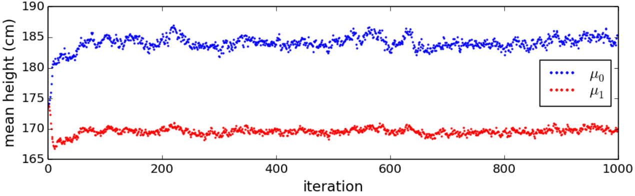
- Even though we're not using the true labels, it is interesting to note that this is fairly close to the sample averages: 168.0 cm (5 feet 6.1 inches) for females, and 181.4 cm (5 feet 11.4 inches) for males.

- The traceplot of π indicates that the sampler is exploring values of around 0.2 to 0.4—that is, the proportion of people coming from group 1 is around 0.2 to 0.4.

- Meanwhile, the true empirical proportion of males is $562/(695 + 562) \approx 0.45$. So the posterior seems slightly off. This could be due to not having enough data, and/or due to the fact that we are assuming a shared, fixed value of λ.

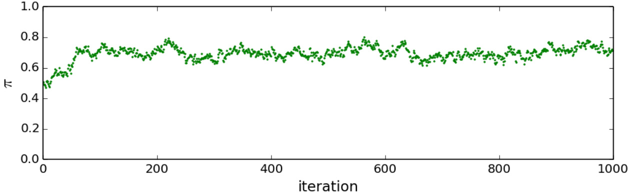
Height example: Results from another run . . . uh oh!

Traceplots of the component means μ_0 and μ_1



Traceplot of the mixture weight π

($\pi =$ prior probability that a subject comes from component 1)



Height example: A potentially serious issue

- Why are females assigned to component 0 and males assigned to component 1? Why not the other way around?
- The model is symmetric with respect to the two components, and thus the posterior is also symmetric.
- If we run the sampler multiple times, it will randomly settle in one of these two modes.

Height example: A potentially serious issue

- If the sampler were behaving properly, it would move back and forth between these two modes, but it doesn't—it gets stuck in one and stays there.
- This is a very common problem with mixture models. Fortunately, however, the results are still valid if we interpret them correctly.
- Specifically, our inferences will be valid as long as we only consider quantities that are invariant with respect to permutations of the components.

Outline

Introduction

K-means clustering

Gaussian mixture model

Bayesian Gaussian mixture model

Gibbs sampler for GMM

Height example

Tricky aspects of Bayesian mixture models

Label switching problem

- The mixture likelihood is invariant to permutations of the component assignment labels.
  - E.g., in the height example, female/male could be 0/1 or 1/0.
- Thus, if the prior is invariant, the posterior is invariant as well.
- This symmetry means that there are typically $K!$ regions of parameter space with high posterior probability.
- Since this is a nonidentifiability of the model, it doesn't really matter which permutation of labels we use.
- However, there is a subtle issue. Suppose MCMC is mixing well enough that it moves between multiple permutations.
  - How would you estimate, say, the posterior mean of the female heights?
  - Consider the MCMC samples of μ_0. What will the sample average converge to? What about μ_1?

Label switching problem

- The *label switching problem* is that MCMC sample averages of permutation-dependent quantities are usually meaningless, if the MCMC chain is mixing between multiple permutations.
- The most obvious "solution" is to impose constraints to ensure identifiability, however, this doesn't always solve the problem.
- The reason is that the constraint boundaries may chop up some of the $K!$ "modes" into two or more parts.

Label switching problem: Galaxy example

Galaxy dataset example

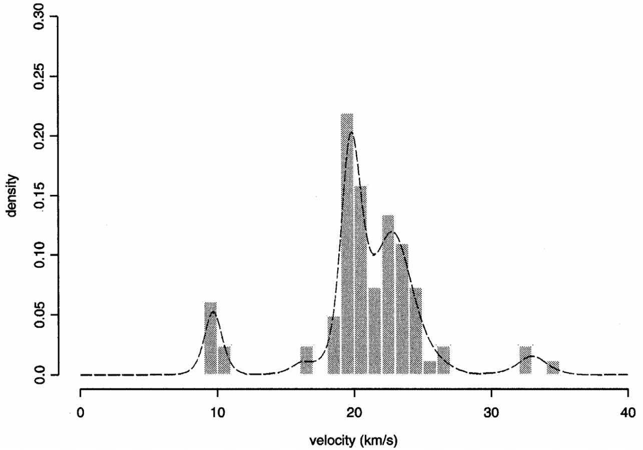


FIG. 3. *Histogram of the Galaxy data. We have overlaid the histogram with a kernel density estimate (dashed).*

(Jasra et al., 2005)

Label switching problem: Galaxy example

Identifiability constraints don't solve the problem

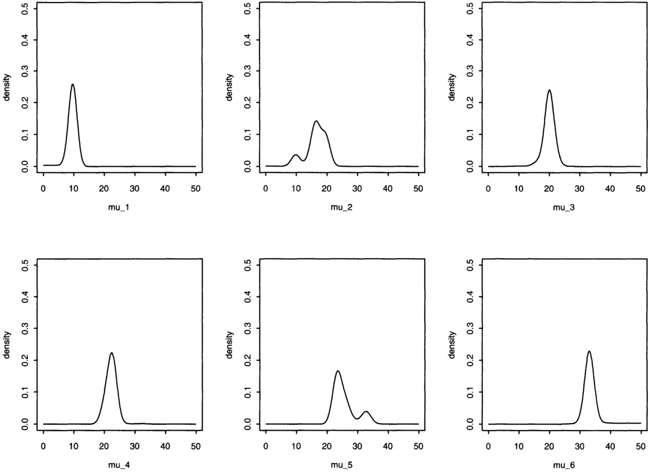


FIG. 4. *Marginal posterior density estimates of the sampled means of the galaxy data set. The means were permuted to obey the constraint* $\mu_1 < \cdots < \mu_6$. *We fitted a six component normal mixture to the data. The output is the last* 20,000 *iterations from the Gibbs sampler.*

(Jasra et al., 2005)

Label switching problem: Possible solutions

- Look at individual samples, e.g., in scatterplots.

- Only take averages of label-invariant quantities.
  - For example, average $I(z_i = z_j)$ to estimate the *similarity matrix* $S_{ij} = \mathbb{P}(Z_i = Z_j | x)$.

- Use label-invariant loss functions to compute posterior summaries (Celeux et al., 2000).

- Relabel each MCMC sample to minimize a loss function that encourages similar points to be together (Stephens, 2000).

- Mean partition: Choose a partition of the datapoints that minimizes distance to the MCMC samples of partitions (Huelsenbeck and Andolfatto, 2007).

- If some labels are available, use them as anchors (Kunkel and Peruggia, 2018).

Don't overinterpret the clusters

- In many applications, a mixture model is used for practical purposes, rather than because the data are actually thought to arise from a mixture.
  - For example, when clustering images or documents.
  - In such cases, one should be careful not to overinterpret the inferred components.

- Meanwhile, sometimes the data definitely come from a mixture, but the assumed model is almost certainly wrong.
  - For example, extracellular recordings of multiple neurons.
  - Again, it is dangerous to overinterpret the inferred components.

- Interpretation of mixture component parameters should only be done with a healthy dose of skepticism.

Choosing the number of components, K

- Choosing K can be tricky.
- Cross-validation is one option, however, it can be computationally expensive.
- The marginal likelihood is not easy to compute, but Pritchard et al. (2000) define an approximate marginal likelihood that is reliable and useful.
- A natural Bayesian approach is to put a prior on K. This works well and is similar to infinite mixture models (Miller and Harrison, 2018).
- A computationally convenient option is to use an “overfitted mixture” in which a large K is used, and the prior on π is chosen to make unneeded components shrink to zero weight.

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
`https://forms.gle/WBzKPGP2XoGtwJcD9`

(Three people per room, randomly assigned. 15 minutes.)

References and supplements: Applications

- Yeung, K. Y., et al. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics, 17(10), 977-987.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.
- Stauffer, C., & Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 246-252.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41.
- Pagel, M., & Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology, 53(4), 571-581.
- Lee, S., & McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: Some recent and new results. Statistics and Computing, 24(2), 181-202.

References and supplements: Methodology

- Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science, 50-67.
- Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association, 95(451), 957-970.
- Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(4), 795-809.
- Huelsenbeck J.P. & Andolfatto P. (2007). Inference of population structure under a Dirichlet process model. Genetics. 175:1787–1802.
- Kunkel, D. & Peruggia, M. (2018). Anchored Bayesian Gaussian Mixture Models. arXiv preprint arXiv:1805.08304.
- Miller, J. W., & Harrison, M. T. (2018). Mixture models with a prior on the number of components. Journal of the American Statistical Association, 113(521), 340-356.