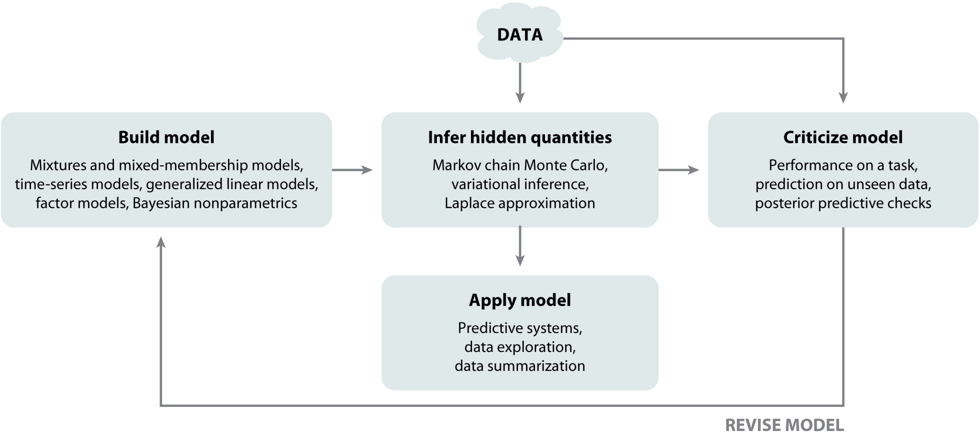
STATS271/371: Applied Bayesian Statistics

Bayesian Mixture Models and (Collapsed) Gibbs Sampling

Scott Linderman

April 21, 2021

Box's Loop



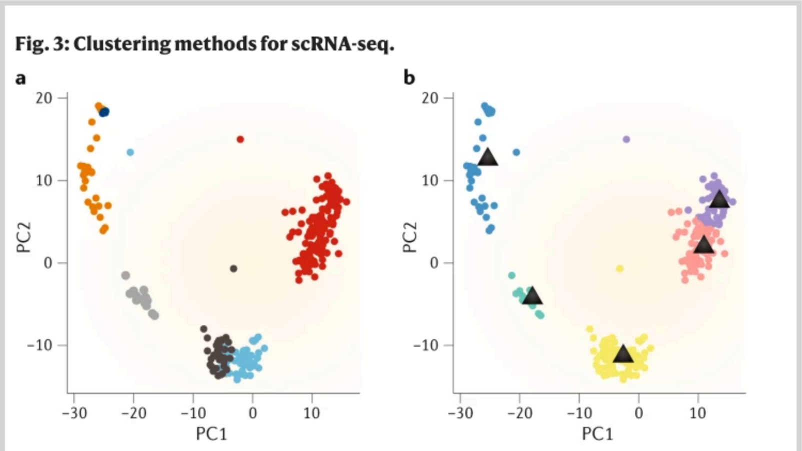
Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

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

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

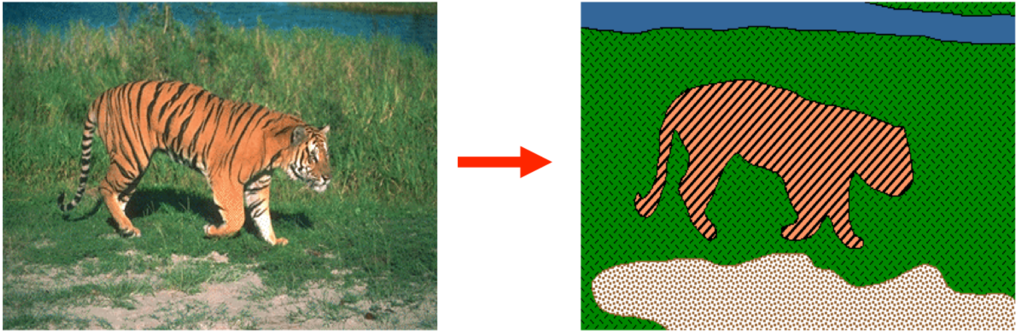
- **Model:** Bayesian mixture models
- **Algorithm:** Gibbs sampling
- **Criticism:** Posterior predictive checks
- **Algorithm II:** Collapsed Gibbs sampling

Motivation: Clustering scRNA-seq data



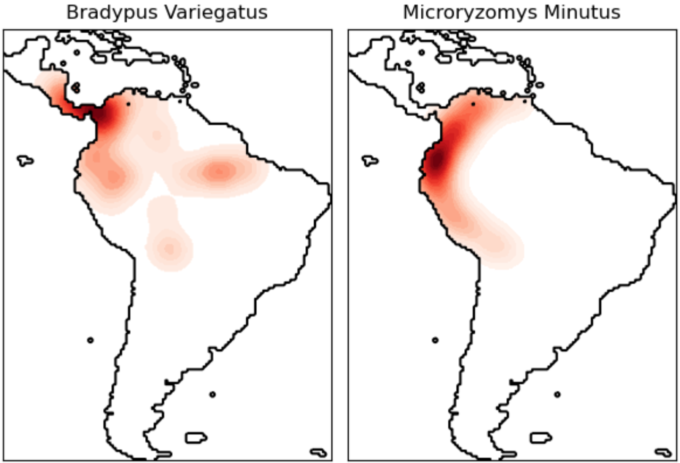
From Kiselev et al. [2019]

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,

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

Equivalently,

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

Substituting in the assumed forms

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

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,

$$\begin{aligned} p(\boldsymbol{\eta}_k \mid -) &\propto \exp\left\{ \left(\boldsymbol{\phi} + \sum_{n=1}^N \mathbb{I}[z_n = k] \boldsymbol{x}_n \right)^\top \boldsymbol{\eta}_k - \frac{\nu + N_k}{2} \boldsymbol{\eta}_k^\top \boldsymbol{\eta}_k \right\} & (27) \\ &\propto \mathcal{N}\left(\boldsymbol{\eta}_k \mid (\nu + N_k)^{-1} \left(\boldsymbol{\phi} + \sum_{n=1}^N \mathbb{I}[z_n = k] \boldsymbol{x}_n \right), (\nu + N_k)^{-1} \boldsymbol{I} \right) & (28) \end{aligned}$$

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

Gibbs sampling in Bayesian 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,

$$p(\boldsymbol{\pi} \mid -) \propto \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n \mid \boldsymbol{\pi}) \tag{30}$$

$$\propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \times \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{\mathbb{I}[z_n = k]} \tag{31}$$

$$\propto \mathrm{Dir}\left(\boldsymbol{\pi} \mid [\alpha_1 + N_1, \ldots, \alpha_K + N_K]\right) \tag{32}$$

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.

Next time

- Show that **Gibbs is a special case of MH** and, as such, asymptotically generates samples from the posterior distribution.
- Talk about **posterior predictive checks** and ways of choosing K.
- Introduce **collapsed Gibbs sampling**, which will allow us to generalize to **nonparametric Bayesian mixture 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.