Bayesian Mixture Models, MAP Estimation, and K-Means

STATS 305C: Applied Statistics

Scott Linderman

April 25, 2023

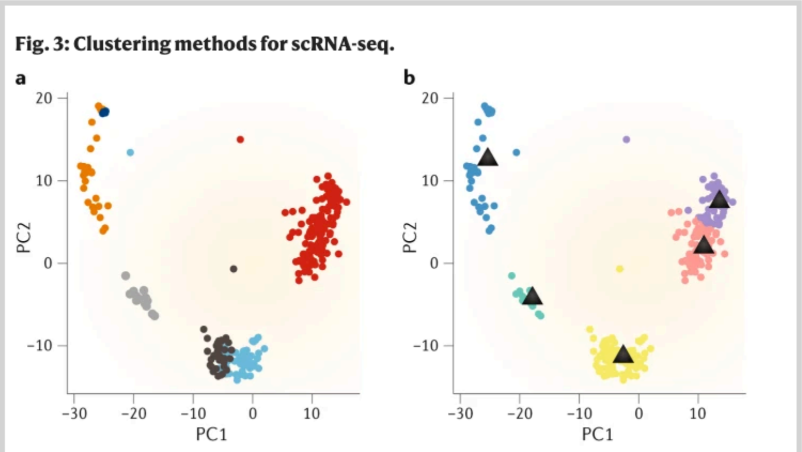
Outline

- Model: Bayesian mixture models
- Algorithm: MAP Estimation / K-Means

Where are we?

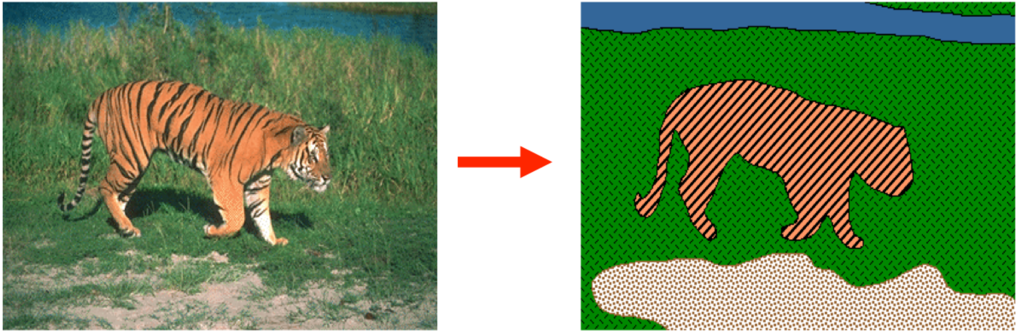
Model	Algorithm	Application
Multivariate Normal Models	Conjugate Inference	Bayesian Linear Regression
Hierarchical Models	MCMC (MH & Gibbs)	Modeling Polling Data
Probabilistic PCA & Factor Analysis	MCMC (HMC)	Images Reconstruction
Mixture Models	EM & Variational Inference	Image Segmentation
Mixed Membership Models	Coordinate Ascent VI	Topic Modeling
Variational Autoencoders	Black Box, Amortized VI	Image Generation
State Space Models	Message Passing	Segmenting Video Data
Bayesian Nonparametrics	Fancy MCMC	Modeling Neural Spike Trains

Motivation: Clustering scRNA-seq data



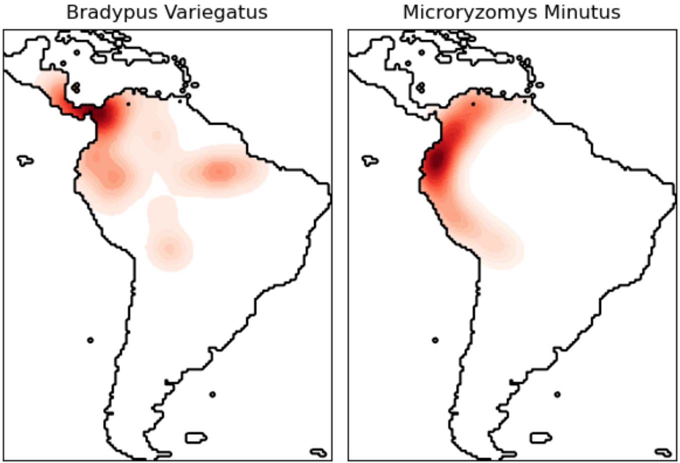
From Kiselev et al. [2019]

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

- $\boldsymbol{\theta}_k$ denote the *natural parameters* of component k
- $\boldsymbol{\pi} \in \Delta_{K-1}$ denote the component *proportions* (i.e. probabilities).

Hyperparameters: Let

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

Generative Model

1. Sample the proportions from a Dirichlet prior:

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

The beta distribution

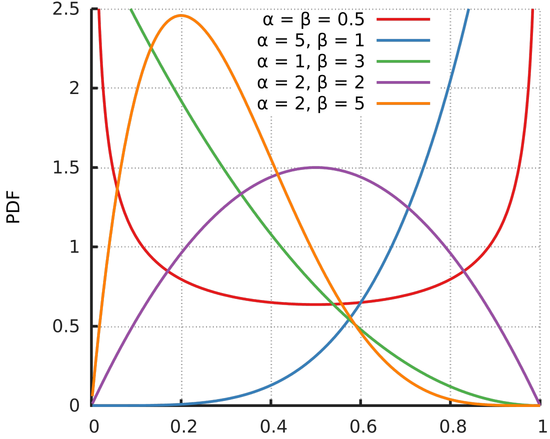


Figure: The beta distribution over $\pi \in [0, 1]$ is a special case of the Dirichlet distribution.
`https://en.wikipedia.org/wiki/Beta_distribution`

The Dirichlet distribution

If the beta distribution generates weighted coins, the Dirichlet generates weighted dice.

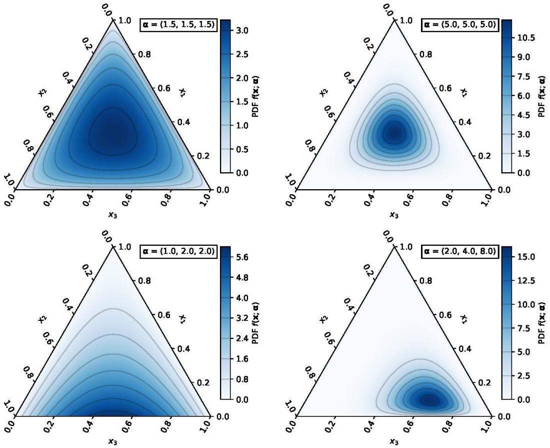


Figure: The Dirichlet distribution over $\pi \in \Delta_2$; i.e. distributions over $K = 3$ outcomes. From `https://en.wikipedia.org/wiki/Dirichlet_distribution`

Generative Model

1. Sample the proportions from a Dirichlet prior:

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{2}$$

2. Sample the parameters for each component:

$$\boldsymbol{\theta}_k \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu) \qquad \text{for } k = 1, \ldots, K \tag{3}$$

3. Sample the assignment of each data point:

$$z_n \overset{\text{iid}}{\sim} \boldsymbol{\pi} \qquad \text{for } n = 1, \ldots, N \tag{4}$$

4. Sample data points given their assignments:

$$\boldsymbol{x}_n \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_{z_n}) \qquad \text{for } n = 1, \ldots, N \tag{5}$$

Joint distribution

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

$$p(\pi, \{\theta_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N p(z_n \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \{\boldsymbol{\theta}_k\}_{k=1}^K) \tag{6}$$

- Equivalently,

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

- Substituting in the assumed forms

$$p(\pi, \{\theta_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K \left[\pi_k\, p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)\right]^{\mathbb{I}[z_n = k]} \tag{8}$$

Joint distribution

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

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N p(z_n \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \{\boldsymbol{\theta}_k\}_{k=1}^K) \tag{6}$$

- Equivalently,

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

- Substituting in the assumed forms

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K [\pi_k\, p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)]^{\mathbb{I}[z_n=k]} \tag{8}$$

Joint distribution

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

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = p(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N p(z_n \mid \pi)\, p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \{\boldsymbol{\theta}_k\}_{k=1}^K) \tag{6}$$

- Equivalently,

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

- Substituting in the assumed forms

$$p(\pi, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) = \mathrm{Dir}(\pi \mid \boldsymbol{\alpha}) \prod_{k=1}^K p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \prod_{n=1}^N \prod_{k=1}^K \left[\pi_k\, p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)\right]^{\mathbb{I}[z_n = k]} \tag{8}$$

Exponential family mixture models

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

Let's assume an **exponential family** likelihood,

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}_k) = h(\boldsymbol{x}_n) \exp\left\{\langle t(\boldsymbol{x}_n), \boldsymbol{\theta}_k\rangle - A(\boldsymbol{\theta}_k)\right\}. \tag{9}$$

Then assume a **conjugate prior**,

$$p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) \propto \exp\left\{\langle \boldsymbol{\phi}, \boldsymbol{\theta}_k\rangle - \nu A(\boldsymbol{\theta}_k)\right\}. \tag{10}$$

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 $\boldsymbol{\theta}_k \in \mathbb{R}^D$ and identity covariance,

$$
\begin{aligned}
p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k) &= \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) && (11)\\
&= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\theta}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\theta}_k)\right\} && (12)\\
&= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{x}_n + \boldsymbol{x}_n^\top \boldsymbol{\theta}_k - \tfrac{1}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right\}, && (13)
\end{aligned}
$$

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 $A(\boldsymbol{\theta}_k) = \frac{1}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k$.

The conjugate prior is a Gaussian prior on the mean,

$$
p(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}, \nu) = \mathcal{N}(\nu^{-1}\boldsymbol{\phi}, \nu^{-1}\boldsymbol{I}) \propto \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \tfrac{\nu}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right\} = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \nu A(\boldsymbol{\theta}_k)\right\}. \qquad (14)
$$

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

Outline

- Model: Bayesian mixture models
- **Algorithm: MAP Estimation / K-Means**

MAP inference via coordinate ascent

Let's first consider **maximum a posteriori (MAP) inference**.

Idea: find the mode of $p(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{z_n\}_{n=1}^N \mid \{\mathbf{x}_n\}_{n=1}^N, \boldsymbol{\phi}, \nu, \boldsymbol{\alpha})$ 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 $\boldsymbol{\pi} = \frac{1}{K}\mathbf{1}_K$.

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

$$p(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto p(\boldsymbol{x}_n \mid z_n, \{\boldsymbol{\theta}_k\}_{k=1}^K) = \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_{z_n}, \boldsymbol{I}) \tag{15}$$

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

$$z_n^\star = \underset{k \in \{1,\ldots,K\}}{\arg\min} \|\boldsymbol{x}_n - \boldsymbol{\theta}_k\|_2. \tag{16}$$

Coordinate ascent in the Gaussian mixture model II

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

$$p(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \{(z_n, \boldsymbol{x}_n)\}_{n=1}^N \mid \boldsymbol{\phi}, \nu, \boldsymbol{\alpha}) \propto \prod_{n=1}^N p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k)^{\mathbb{I}[z_n=k]} \tag{17}$$

$$\propto \exp\left\{\sum_{n=1}^N \mathbb{I}[z_n = k]\left(\boldsymbol{x}_n^\top \boldsymbol{\theta}_k - \tfrac{1}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right)\right\} \tag{18}$$

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

$$\boldsymbol{\theta}_k^\star = \frac{1}{N_k}\sum_{n=1}^K \mathbb{I}[z_n = k]\boldsymbol{x}_n, \tag{19}$$

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

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

EM in the Gaussian mixture model

K-Means made **hard assignments** of data points to clusters in each iteration. What if we used **soft assignments** instead?

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,

$$\omega_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_j, \boldsymbol{I})}. \tag{20}$$

2. For each component k, set the new mean to

$$\boldsymbol{\theta}_k^\star = \frac{1}{N_k} \sum_{n=1}^{K} \omega_{nk} \boldsymbol{x}_n, \tag{21}$$

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

This is called the **expectation maximization (EM)** algorithm.

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.