Bayesian Mixture Models and Expectation Maximization

STATS 305C: Applied Statistics

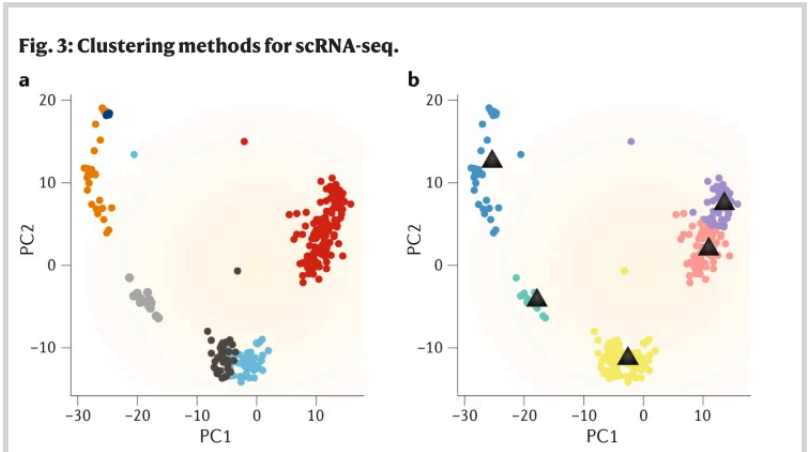
Scott Linderman

April 18, 2022

Outline

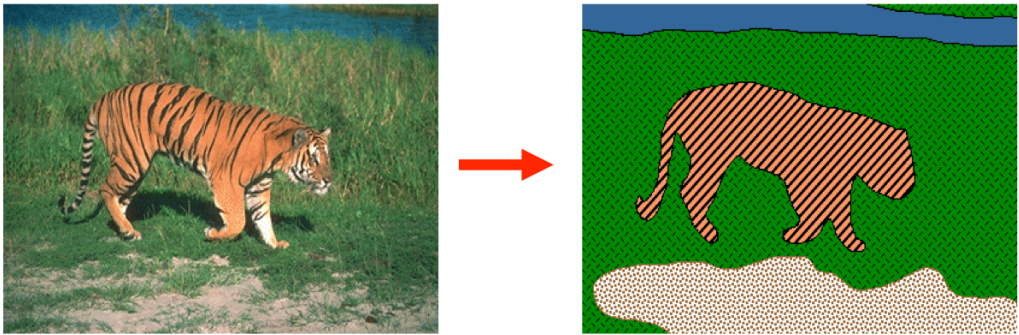
- Model: Bayesian mixture models
- Algorithm: MAP Estimation / K-Means
- Algorithm: Expectation Maximization

Motivation: Clustering scRNA-seq data



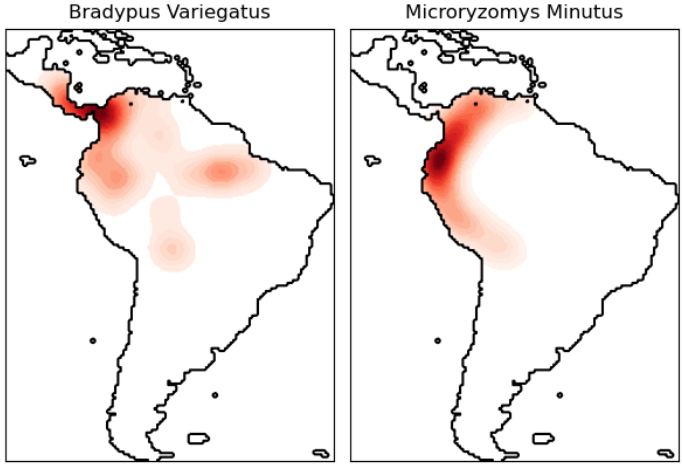
From Kiselev et al. [2019]

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:

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

The beta distribution

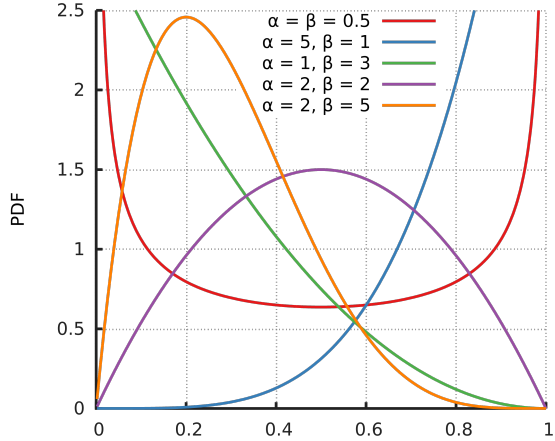


Figure: The beta distribution over $\pi \in [0, 1]$ is a special case of the Dirichlet distribution.
`https://en.wikipedia.org/wiki/Beta_distribution`

The Dirichlet distribution

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

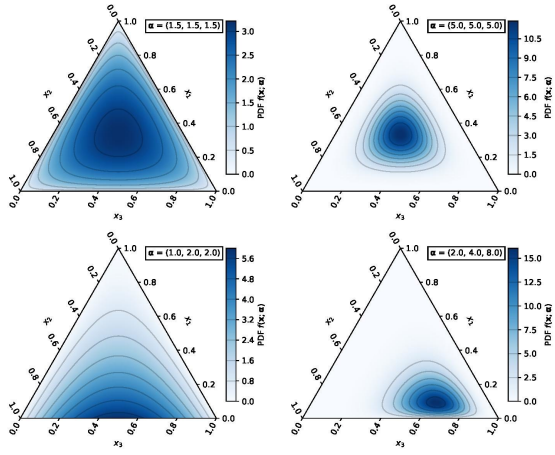


Figure: The Dirichlet distribution over $\pi \in \Delta_2$; i.e. distributions over $K = 3$ outcomes. From `https://en.wikipedia.org/wiki/Dirichlet_distribution`

Generative Model

1. Sample the proportions from a Dirichlet prior:

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

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:

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

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,

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

- 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,

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

- 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,

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

- Substituting in the assumed forms

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

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,

$$p(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) \tag{11}$$

$$= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\theta}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\theta}_k)\right\} \tag{12}$$

$$= (2\pi)^{-D/2} \exp\left\{-\tfrac{1}{2}\boldsymbol{x}_n^\top \boldsymbol{x}_n + \boldsymbol{x}_n^\top \boldsymbol{\theta}_k - \tfrac{1}{2}\boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k\right\}, \tag{13}$$

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,

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

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**
- Algorithm: Expectation Maximization

MAP inference via coordinate ascent

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

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

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

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

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**!

Outline

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

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.

What is EM doing?

Rather than maximizing the **joint probability**, EM is maximizing the **marginal probability**,

$$\log p(\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \tag{22}$$

$$= \log p(\boldsymbol{\theta}) + \log \prod_{n=1}^{N} \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) \tag{23}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) \tag{24}$$

For discrete mixtures (with small enough K) we can evaluate the log marginal probability (with what complexity?).

We can usually evaluate its gradient too, so we could just do gradient ascent to find $\boldsymbol{\theta}^*$.

However, EM typically obtains faster convergence rates.

What is EM doing? II

Idea: Obtain a lower bound on the marginal probability,

$$\log p(\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \sum_{z_n} p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) \tag{25}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \sum_{z_n} q(z_n) \frac{p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}{q(z_n)} \tag{26}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \mathbb{E}_{q(z_n)} \left[\frac{p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}{q(z_n)} \right] \tag{27}$$

where $q(z_n)$ is any distribution on $z_n \in \{1, \ldots, K\}$ such that $p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})$ is **absolutely continuous** w.r.t. $q(z_n)$.

Jensen's Inequality

Jensen's inequality states that,

$$f(\mathbb{E}_{p(y)}[y]) \geq \mathbb{E}_{p(y)}[f(y)] \tag{28}$$

if f is a **concave function**, with equality iff f is linear.

[Picture]

What is EM doing? III

Applied to the log marginal probability, Jensen's inequality yields,

$$\log p(\boldsymbol{X}, \boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log \mathbb{E}_{q_n(z_n)} \left[\frac{p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})}{q_n(z_n)} \right] \tag{29}$$

$$\geq \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} \left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n) \right] \tag{30}$$

$$\triangleq \mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] \tag{31}$$

where $\boldsymbol{q} = (q_1, \ldots, q_N)$ is a tuple of densities.

This is called the **evidence lower bound**, or **ELBO** for short.

It is a function of $\boldsymbol{\theta}$ and a **functional** of $\boldsymbol{q}$, since each q_n is a probability density function.

We can think of **EM as coordinate ascent on the ELBO**.

M-step: Maximizing the ELBO wrt θ (Gaussian case)

Suppose we fix $\boldsymbol{q}$. Since each z_n is a discrete latent variable, q_n must be a probability mass function. Let it be denoted by,

$$q_n(z_n) = [q_n(z_n = 1), \ldots, q_n(z_n = K)]^\top = [\omega_{n1}, \ldots, \omega_{nK}]^\top. \tag{32}$$

(These will be the **responsibilities** from before.)

Now, recall our basic model, $\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\theta}_{z_n}, \boldsymbol{I})$, and assume a prior $\boldsymbol{\theta}_k \sim \mathcal{N}(\boldsymbol{\phi}, \nu^{-1}\boldsymbol{I})$, Then,

$$\mathcal{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)}[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})] + c \tag{33}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \log p(\boldsymbol{x}_n, z_n = k \mid \boldsymbol{\theta}) + c \tag{34}$$

$$= \sum_{k=1}^{K} \left[\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \tfrac{\nu}{2} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k \right] + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \left[\boldsymbol{x}_n^\top \boldsymbol{\theta}_k - \tfrac{1}{2} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k \right] + c \tag{35}$$

M-step: Maximizing the ELBO wrt θ (Gaussian case) II

Zooming in on just $\boldsymbol{\theta}_k$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \boldsymbol{\phi}_{N,k}^\top \boldsymbol{\theta}_k - \tfrac{1}{2} \nu_{N,k} \boldsymbol{\theta}_k^\top \boldsymbol{\theta}_k \tag{36}$$

where

$$\boldsymbol{\phi}_{N,k} = \boldsymbol{\phi} + \sum_{n=1}^{N} \omega_{nk} \boldsymbol{x}_n \qquad \nu_{N,k} = \nu + \sum_{n=1}^{N} \omega_{nk} \tag{37}$$

Taking derivatives and setting to zero yields,

$$\boldsymbol{\theta}_k^\star = \frac{\boldsymbol{\phi}_{N,k}}{\nu_{N,k}} = \frac{\boldsymbol{\phi} + \sum_{n=1}^{N} \omega_{nk} \boldsymbol{x}_n}{\nu + \sum_{n=1}^{N} \omega_{nk}}. \tag{38}$$

In the improper uniform prior limit where $\boldsymbol{\phi} \to 0$ and $\nu \to 0$, we recover the EM updates shown on slide 20.

E-step: Maximizing the ELBO wrt q (Gaussian case)

As a function of q_n, for discrete Gaussian mixtures with identity covariance,

$$\mathcal{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \mathbb{E}_{q_n(z_n)}\left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)\right] + c \tag{39}$$

$$= \sum_{k=1}^{K} \omega_{nk}\left[\log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) + \log \pi_k - \log \omega_{nk}\right] + c \tag{40}$$

where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^\top$ is the vector of cluster probabilities.

We also have two constraints: $\omega_{nk} \geq 0$ and $\sum_k \omega_{nk} = 1$. Let's ignore the non-negative constraint for now (it will automatically be satisfied anyway) and write the Lagrangian with the simplex constraint,

$$\mathcal{J}(\boldsymbol{\omega}_n, \lambda) = \sum_{k=1}^{K} \omega_{nk}\left[\log \mathcal{N}(\boldsymbol{x}_n \mid \theta_k, \boldsymbol{I}) + \log \pi_k - \log \omega_{nk}\right] - \lambda\left(1 - \sum_{k=1}^{K} \omega_{nk}\right) \tag{41}$$

E-step: Maximizing the ELBO wrt q (Gaussian case) II

Taking the partial derivative wrt ω_{nk} and setting to zero yields,

$$\frac{\partial}{\partial \omega_{nk}} \mathcal{J}(\boldsymbol{\omega}_n, \lambda) = \log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) + \log \pi_k - \log \omega_{nk} - 1 + \lambda = 0 \tag{42}$$

$$\Rightarrow \log \omega_{nk}^\star = \log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) + \log \pi_k + \lambda - 1 \tag{43}$$

$$\Rightarrow \omega_{nk}^\star \propto \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I}) \tag{44}$$

Enforcing the simplex constraint yields,

$$\omega_{nk}^\star = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_k, \boldsymbol{I})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\theta}_j, \boldsymbol{I})}, \tag{45}$$

just like on slide 20.

Note that

$$\omega_{nk}^\star \propto p(z_n = k)\, p(\boldsymbol{x}_n \mid z_n = k, \boldsymbol{\theta}) = p(z_n = k \mid \boldsymbol{x}_n, \boldsymbol{\theta}) \tag{46}$$

The ELBO is tight after the E-step

Equivalently, q_n equals the posterior, $p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$. At that point, the ELBO simplifies to,

$$\mathcal{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} \left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)\right] \tag{47}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})} \left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})\right] \tag{48}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})} \left[\log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})\right] \tag{49}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) \tag{50}$$

$$= \log p(\boldsymbol{X}, \boldsymbol{\theta}) \tag{51}$$

In other words, **after the E step, the bound is tight**!

EM as a minorize-maximize (MM) algorithm

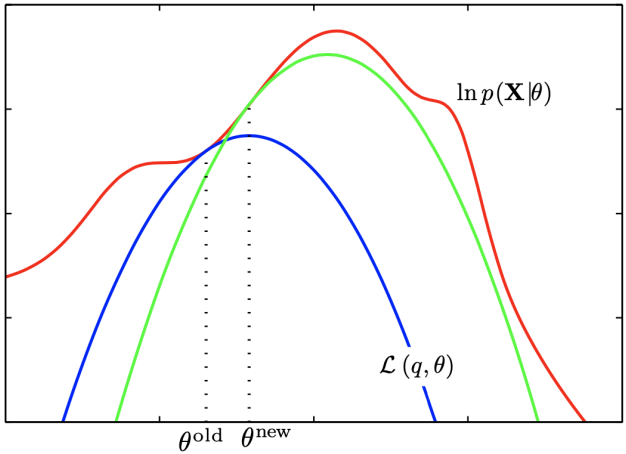


Figure: Bishop, Figure 9.14: EM alternates between constructing a lower bound (minorizing) and finding new parameters that maximize it.

M-step: Maximizing the ELBO wrt θ (generic exp. fam.)

Now let's consider the general Bayesian mixture with exponential family likelihoods and conjugate priors. As a function of $\boldsymbol{\theta}$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)}[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta})] + \text{c} \tag{52}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \log p(\boldsymbol{x}_n, z_n = k \mid \boldsymbol{\theta}) + \text{c} \tag{53}$$

$$= \sum_{k=1}^{K} \left[\boldsymbol{\phi}^\top \boldsymbol{\theta}_k - \nu A(\boldsymbol{\theta}_k) \right] + \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk} \left[t(\boldsymbol{x}_n)^\top \boldsymbol{\theta}_k - A(\boldsymbol{\theta}_k) \right] + \text{c} \tag{54}$$

M-step: Maximizing the ELBO wrt θ (generic exp. fam.) II

Zooming in on just θ_k,

$$\mathscr{L}[\theta, q] = \phi_{N,k}^\top \theta_k - \nu_{N,k} A(\theta_k) \tag{55}$$

where

$$\phi_{N,k} = \phi + \sum_{n=1}^{N} \omega_{nk} t(x_n) \qquad \nu_{N,k} = \nu + \sum_{n=1}^{N} \omega_{nk} \tag{56}$$

Taking derivatives and setting to zero yields,

$$\theta_k^* = \left[\frac{\mathrm{d}}{\mathrm{d}\theta_k} A\right]^{-1} \left(\frac{\phi_{N,k}}{\nu_{N,k}}\right) \tag{57}$$

TODO: Derivatives of log normalizers, expected sufficient statistics, ...

E-step: Maximizing the ELBO wrt q (generic exp. fam.)

In our first pass, we used the fact that q_n is a finite pmf in the discrete mixture model case, but now let's take a more general perspective.

More generally, q_n will be a probability density function, and optimizing over functions requires the **calculus of variations**.

However, note that we can write the ELBO in a slightly different form,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} \left[\log p(\boldsymbol{x}_n, z_n \mid \boldsymbol{\theta}) - \log q_n(z_n)\right] \tag{58}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \mathbb{E}_{q_n(z_n)} \left[\log p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) + \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) - \log q_n(z_n)\right] \tag{59}$$

$$= \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} \left[\log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) - D_{\mathrm{KL}}\left(q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})\right)\right] \tag{60}$$

where $D_{\mathrm{KL}}\left(\cdot \,\|\, \cdot\right)$ denote the **Kullback-Leibler divergence**.

Kullback-Leibler (KL) divergence

The KL divergence is defined as,

$$D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) = \int q(z) \log \frac{q(z)}{p(z)} \,\mathrm{d}z. \tag{61}$$

It gives a notion of how similar two distributions are, but it is **not a distance!** (It is not symmetric, e.g.) Still, it has some intuitive properties:

- It is non-negative, $D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) \geq 0$.
- It equals zero iff the distributions are the same, $D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) = 0 \iff q(z) = p(z)$ almost everywhere.

E-step: Maximizing the ELBO wrt q (generic exp. fam.) II

Maximizing the ELBO wrt q_n amounts to minimizing the KL divergence to the posterior $p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$,

$$\mathscr{L}[\boldsymbol{\theta}, \boldsymbol{q}] = \log p(\boldsymbol{\theta}) + \sum_{n=1}^{N} [\log p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) - D_{\mathrm{KL}}(q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}))] \tag{62}$$

$$= -D_{\mathrm{KL}}(q_n(z_n) \,\|\, p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})) + c \tag{63}$$

As we said, the KL is minimized when $q_n(z_n) = p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$, so the optimal update is,

$$q_n^\star(z_n) = p(z_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}), \tag{64}$$

just like we found on slide 28.

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.