Machine Learning for OR & FE

The EM Algorithm

Martin Haugh

Department of Industrial Engineering and Operations Research
Columbia University
Email: martin.b.haugh@gmail.com

Outline

The EM Algorithm
E.G. Missing Data in a Multinomial Model
E.G. Normal Mixture Models Revisited

Detour: Kullback-Leibler Divergence
The EM Algorithm Revisited
E.G. Questionnaires and Missing Data

The EM Algorithm (for Computing ML Estimates)

Assume the complete data-set consists of $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$

– but only $\mathcal{X}$ is observed.

The complete-data log likelihood is denoted by $l(\theta; \mathcal{X}, \mathcal{Y})$ where θ is the unknown parameter vector for which we wish to find the MLE.

E-Step: Compute the expected value of $l(\theta; \mathcal{X}, \mathcal{Y})$ given the observed data, $\mathcal{X}$, and the current parameter estimate θ_{old}. In particular, we define

$$\begin{aligned} Q(\theta; \theta_{old}) &:= \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old} \right] \\ &= \int l(\theta; \mathcal{X}, y)\, p(y \mid \mathcal{X}, \theta_{old})\, dy \end{aligned} \tag{1}$$

– $p(\cdot \mid \mathcal{X}, \theta_{old}) \equiv$ conditional density of $\mathcal{Y}$ given observed data, $\mathcal{X}$, and θ_{old}

– $Q(\theta; \theta_{old})$ is the expected complete-data log-likelihood.

M-Step: Compute $\theta_{new} := \max_\theta \; Q(\theta; \theta_{old})$.

The EM Algorithm

Now set $\theta_{old} = \theta_{new}$ and iterate E- and M-steps until sequence of θ_{new}'s converges.

Convergence to a local maximum can be guaranteed under very general conditions

– will see why below.

If suspected that log-likelihood function has multiple local maximums then the EM algorithm should be run many times

– using a different starting value of θ_{old} on each occasion.

The ML estimate of θ is then taken to be the best of the set of local maximums obtained from the various runs of the EM algorithm.

Why Does the EM Algorithm Work?

Will use $p(\cdot \mid \cdot)$ to denote a generic conditional PDF. Now observe that

$$
\begin{aligned}
l(\theta; \mathcal{X}) &= \ln p(\mathcal{X} \mid \theta) \\
&= \ln \int p(\mathcal{X}, y \mid \theta)\ dy \\
&= \ln \int \frac{p(\mathcal{X}, y \mid \theta)}{p(y \mid \mathcal{X}, \theta_{old})} p(y \mid \mathcal{X}, \theta_{old})\ dy \\
&= \ln \mathsf{E}\left[\frac{p(\mathcal{X}, \mathcal{Y} \mid \theta)}{p(\mathcal{Y} \mid \mathcal{X}, \theta_{old})} \mid \mathcal{X}, \theta_{old}\right] \\
&\geq \mathsf{E}\left[\ln\left(\frac{p(\mathcal{X}, \mathcal{Y} \mid \theta)}{p(\mathcal{Y} \mid \mathcal{X}, \theta_{old})}\right) \mid \mathcal{X}, \theta_{old}\right] \quad \text{by Jensen's inequality} \qquad (2) \\
&= \mathsf{E}\left[\ln p(\mathcal{X}, \mathcal{Y} \mid \theta) \mid \mathcal{X}, \theta_{old}\right] \ - \ \mathsf{E}\left[\ln p(\mathcal{Y} \mid \mathcal{X}, \theta_{old}) \mid \mathcal{X}, \theta_{old}\right] \\
&= Q(\theta; \theta_{old}) \ - \ \mathsf{E}\left[\ln p(\mathcal{Y} \mid \mathcal{X}, \theta_{old}) \mid \mathcal{X}, \theta_{old}\right] \qquad (3)
\end{aligned}
$$

Also clear (why?) that inequality in (2) is an equality if we take $\theta = \theta_{old}$.

Why Does the EM Algorithm Work?

Let $g(\theta \mid \theta_{old})$ denote the right-hand-side of (3).

Therefore have

$$l(\theta; \mathcal{X}) \;\geq\; g(\theta \mid \theta_{old})$$

for all θ with equality when $\theta = \theta_{old}$.

So any value of θ that increases $g(\theta \mid \theta_{old})$ beyond $g(\theta_{old} \mid \theta_{old})$ must also increase $l(\theta; \mathcal{X})$ beyond $l(\theta_{old}; \mathcal{X})$.

The M-step finds such a θ by maximizing $Q(\theta; \theta_{old})$ over θ

– this is equivalent (why?) to maximizing $g(\theta \mid \theta_{old})$ over θ.

Also worth noting that in many applications the function $Q(\theta; \theta_{old})$ will be a convex function of θ

– and therefore easy to optimize.

Schematic for general E-M algorithm

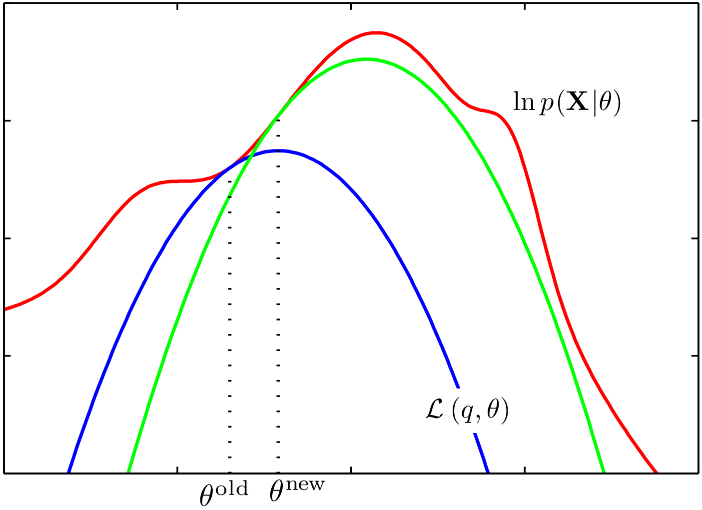

Figure 9.14 from Bishop (where $\mathcal{L}(q, \theta)$ is $g(\theta \mid \theta_{old})$ in our notation)

E.G. Missing Data in a Multinomial Model

Suppose $\mathbf{x} := (x_1, x_2, x_3, x_4)$ is a sample from a $\mathrm{Mult}(n, \pi_\theta)$ distribution where

$$\pi_\theta = \left(\frac{1}{2} + \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

The likelihood, $L(\theta; \mathbf{x})$, is then given by

$$L(\theta; \mathbf{x}) = \frac{n!}{x_1!x_2!x_3!x_4!} \left(\frac{1}{2} + \frac{1}{4}\theta\right)^{x_1} \left(\frac{1}{4}(1-\theta)\right)^{x_2} \left(\frac{1}{4}(1-\theta)\right)^{x_3} \left(\frac{1}{4}\theta\right)^{x_4}$$

so that the log-likelihood $l(\theta; \mathbf{x})$ is

$$l(\theta; \mathbf{x}) = C + x_1 \ln\left(\frac{1}{2} + \frac{1}{4}\theta\right) + (x_2 + x_3)\ln(1-\theta) + x_4 \ln(\theta)$$

– where C is a constant that does not depend on θ.

Could try to maximize $l(\theta; \mathbf{x})$ over θ directly using standard non-linear optimization algorithms

– but we will use the EM algorithm instead.

E.G. Missing Data in a Multinomial Model

To do this we assume the complete data is given by $\mathbf{y} := (y_1, y_2, y_3, y_4, y_5)$ and that $\mathbf{y}$ has a Mult(n, π_θ^*) distribution where

$$\pi_\theta^* = \left(\frac{1}{2},\ \frac{1}{4}\theta,\ \frac{1}{4}(1-\theta),\ \frac{1}{4}(1-\theta),\ \frac{1}{4}\theta\right).$$

However, instead of observing $\mathbf{y}$ we only observe $(y_1 + y_2, y_3, y_4, y_5)$, i.e, $\mathbf{x}$.

Therefore take $\mathcal{X} = (y_1 + y_2, y_3, y_4, y_5)$ and take $\mathcal{Y} = y_2$.

Log-likelihood of complete data then given by

$$l(\theta; \mathcal{X}, \mathcal{Y}) = C + y_2 \ln(\theta) + (y_3 + y_4) \ln(1-\theta) + y_5 \ln(\theta)$$

where again C is a constant containing all terms that do not depend on θ.

Also "clear" that conditional "density" of $\mathcal{Y}$ satisfies

$$f(\mathcal{Y} \mid \mathcal{X}, \theta) = \text{Bin}\left(y_1 + y_2,\ \frac{\theta/4}{1/2 + \theta/4}\right).$$

Can now implement the E-step and M-step.

E.G. Missing Data in a Multinomial Model

Recall that $Q(\theta; \theta_{old}) := \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}\right]$.

E-Step: Therefore have

$$\begin{aligned} Q(\theta; \theta_{old}) &:= C \;+\; \mathsf{E}\left[y_2 \ln(\theta) \;\middle|\; \mathcal{X}, \theta_{old}\right] \;+\; (y_3 + y_4)\ln(1-\theta) \;+\; y_5 \ln(\theta) \\ &= C \;+\; (y_1 + y_2) p_{old} \ln(\theta) \;+\; (y_3 + y_4)\ln(1-\theta) \;+\; y_5 \ln(\theta) \end{aligned}$$

where

$$p_{old} := \frac{\theta_{old}/4}{1/2 + \theta_{old}/4}. \tag{4}$$

E.G. Missing Data in a Multinomial Model

M-Step: Must now maximize $Q(\theta; \theta_{old})$ to find θ_{new}.

Taking the derivative we obtain

$$\begin{aligned} \frac{dQ}{d\theta} &= \frac{(y_1 + y_2)}{\theta} p_{old} - \frac{(y_3 + y_4)}{1 - \theta} + \frac{y_5}{\theta} \\ &= 0 \quad \text{when } \theta = \theta_{new} \end{aligned}$$

where

$$\theta_{new} := \frac{y_5 + p_{old}(y_1 + y_2)}{y_3 + y_4 + y_5 + p_{old}(y_1 + y_2)}. \tag{5}$$

Equations (4) and (5) now define the EM iteration

– which begins with some judiciously chosen value of θ_{old}.

E.G. Normal Mixture Models Revisited

Clustering via normal mixture models is an example of probabilistic clustering

– we assume the data are IID draws

– will consider only the scalar case but note the vector case is similar.

So suppose $\mathcal{X} = (X_1, \ldots, X_n)$ are IID random variables each with PDF

$$f_x(x) \;=\; \sum_{j=1}^{m} p_j \, \frac{e^{-(x-\mu_j)^2/2\sigma_j^2}}{\sqrt{2\pi\sigma_j^2}}$$

where $p_j \geq 0$ for all j and where $\sum_j p_j = 1$

– parameters are the p_j's, μ_j's and σ_j's

– typically estimated via MLE

– which we can do via the EM algorithm.

Normal Mixture Models Revisited

We assume the presence of an additional or latent random variable, Y, where

$$P(Y = j) = p_j, \quad j = 1, \ldots, m.$$

Realized value of Y then determines which of the m normals generates the corresponding value of X

– so there are n such random variables, $(Y_1, \ldots, Y_n) := \mathcal{Y}$.

Note that

$$f_{x|y}(x_i \mid y_i = j, \theta) \;=\; \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(x_i-\mu_j)^2/2\sigma_j^2} \tag{6}$$

where $\theta := (p_1, \ldots, p_m, \mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m)$ is the unknown parameter vector.

The complete data likelihood is given by

$$L(\theta; \mathcal{X}, \mathcal{Y}) = \prod_{i=1}^{n} p_{y_i} \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} e^{-(x_i-\mu_{y_i})^2/2\sigma_{y_i}^2}.$$

Normal Mixture Models

The EM algorithm starts with an initial guess, θ_{old}, and then iterates the E-step and M-step until convergence.

E-Step: Need to compute $Q(\theta;\theta_{old}) := \mathsf{E}\left[l(\theta;\mathcal{X},\mathcal{Y}) \mid \mathcal{X},\theta_{old}\right]$.

Straightforward to show that

$$Q(\theta;\theta_{old}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m} P(\,Y_i = j \mid x_i,\theta_{old})\;\ln\left(f_{x|y}(x_i \mid y_i = j,\theta)\,P(\,Y_i = j \mid \theta)\right). \tag{7}$$

Note that $f_{x|y}(x_i \mid y_i = j,\theta)$ is given by (6) and that $P(\,Y_i = j \mid \theta_{old}) = p_{j,old}$.

Finally, can compute (7) analytically since

$$\begin{aligned} P(\,Y_i = j \mid x_i,\theta_{old}) &= \frac{P(\,Y_i = j,\; X_i = x_i \mid \theta_{old})}{P(X_i = x_i \mid \theta_{old})} \\ &= \frac{f_{x|y}(x_i \mid y_i = j,\theta_{old})\;P(\,Y_i = j \mid \theta_{old})}{\sum_{k=1}^{m} f_{x|y}(x_i \mid y_i = k,\theta_{old})\;P(\,Y_i = k \mid \theta_{old})}. \end{aligned} \tag{8}$$

Normal Mixture Models

M-Step: Can now maximize $Q(\theta; \theta_{old})$ by setting the vector of partial derivatives, $\partial Q / \partial \theta$, equal to 0 and solving for θ_{new}.

After some algebra, we obtain

$$\mu_{j,new} = \frac{\sum_{i=1}^{n} x_i P(Y_i = j \mid x_i, \theta_{old})}{\sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old})} \tag{9}$$

$$\sigma^2_{j,new} = \frac{\sum_{i=1}^{n} (x_i - \mu_j)^2 P(Y_i = j \mid x_i, \theta_{old})}{\sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old})} \tag{10}$$

$$p_{j,new} = \frac{1}{n} \sum_{i=1}^{n} P(Y_i = j \mid x_i, \theta_{old}). \tag{11}$$

Given an initial estimate, θ_{old}, the EM algorithm cycles through (9) to (11) repeatedly, setting $\theta_{old} = \theta_{new}$ after each cycle, until the estimates converge.

Kullback-Leibler Divergence

Let P and Q be two probability distributions such that if $Q(\mathbf{x}) = 0$ then $P(\mathbf{x}) = 0$.

The Kullback-Leibler (KL) divergence or relative entropy of Q from P is defined to be

$$\mathsf{KL}(P \,||\, Q) \;=\; \int_{\mathbf{x}} P(\mathbf{x}) \ln \left(\frac{P(\mathbf{x})}{Q(\mathbf{x})} \right) \qquad (12)$$

with the understanding that $0 \log 0 = 0$.

The KL divergence is a fundamental concept in information theory and machine learning.

Can imagine P representing some true but unknown distribution that we approximate with Q

- $\mathsf{KL}(P \,||\, Q)$ then measures the "distance" between P and Q.

This interpretation is valid because we will see below that $\mathsf{KL}(P \,||\, Q) \geq 0$

- with equality if and only if $P = Q$.

Kullback-Leibler Divergence

The KL divergence is not a true measure of distance since it is:

1. Asymmetric in that $\text{KL}(P \,||\, Q) \neq \text{KL}(Q \,||\, P)$
2. And does not satisfy the triangle inequality.

In order to see that $\text{KL}(P \,||\, Q) \geq 0$ we first recall that a function $f(\cdot)$ is convex on $\mathbb{R}$ if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \text{for all } \alpha \in [0,1].$$

We also recall Jensen's inequality:

Jensen's Inequality: Let $f(\cdot)$ be a convex function on $\mathbb{R}$ and suppose $\mathsf{E}[X] < \infty$ and $\mathsf{E}[f(X)] < \infty$. Then $f(\mathsf{E}[X]) \leq \mathsf{E}[f(X)]$.

Kullback-Leibler Divergence

Noting that $-\ln(x)$ is a convex function we have

$$
\begin{aligned}
\mathsf{KL}(P \,||\, Q) &= -\int_{\mathbf{x}} P(\mathbf{x}) \ln\left(\frac{Q(\mathbf{x})}{P(\mathbf{x})}\right) \\
&\geq -\ln\left(\int_{\mathbf{x}} P(\mathbf{x})\, \frac{Q(\mathbf{x})}{P(\mathbf{x})}\right) \qquad \text{by Jensen's inequality} \\
&= 0.
\end{aligned}
$$

Moreover it is clear from (12) that $\mathsf{KL}(P \,||\, Q) = 0$ if $P = Q$.

In fact because $-\ln(x)$ is strictly convex it is easy to see that $\mathsf{KL}(P \,||\, Q) = 0$ only if $P = Q$.

A Nice Optimization "Trick"

Suppose $\mathbf{c} \in \mathbb{R}_+^n$ and we wish to maximize $\sum_{i=1}^n c_i \ln(q_i)$ over pmf's, $\mathbf{q} = \{q_1, \ldots, q_n\}$.

Let $\mathbf{p} = \{p_1, \ldots, p_n\}$ where $p_i := c_i / \sum_j c_j$ so that $\mathbf{p}$ is a (known) pmf.

We then have:

$$
\begin{aligned}
\max_{\mathbf{q}} \sum_{i=1}^n c_i \ln(q_i) &= \left(\sum_{i=1}^n c_i\right) \max_{\mathbf{q}} \left\{ \sum_{i=1}^n p_i \ln(q_i) \right\} \\
&= \left(\sum_{i=1}^n c_i\right) \max_{\mathbf{q}} \left\{ \sum_{i=1}^n p_i \ln(p_i) - \sum_{i=1}^n p_i \ln\left(\frac{p_i}{q_i}\right) \right\}, \\
&= \left(\sum_{i=1}^n c_i\right) \left(\sum_{i=1}^n p_i \ln(p_i) - \min_{\mathbf{q}} \mathsf{KL}(\mathbf{p} \,||\, \mathbf{q}) \right)
\end{aligned}
$$

from which it follows (why?) that the optimal $\mathbf{q}^*$ satisfies $\mathbf{q}^* = \mathbf{p}$.

Could have saved some time using this trick in earlier multinomial model example

– in particular obtaining (5)

The EM Algorithm Revisited

As before, goal is to maximize the likelihood function $L(\theta; \mathcal{X})$ which is given by

$$L(\theta; \mathcal{X}) = p(\mathcal{X} \,|\, \theta) = \int_y p(\mathcal{X}, y \,|\, \theta)\, dy. \tag{13}$$

Implicit assumption underlying EM algorithm: it is difficult to optimize $p(\mathcal{X} \,|\, \theta)$ with respect to θ directly

– but much easier to optimize $p(\mathcal{X}, \mathcal{Y} \,|\, \theta)$.

First introduce an arbitrary distribution, $q(\mathcal{Y})$, over the latent variables, $\mathcal{Y}$.

Note we can decompose log-likelihood, $l(\theta; \mathcal{X})$, into two terms according to

$$l(\theta; \mathcal{X}) := \ln p(\mathcal{X} \,|\, \theta) \;=\; \underbrace{\mathcal{L}(q, \theta)}_{\text{“energy”}} \;+\; \mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) \tag{14}$$

The EM Algorithm Revisited

$\mathcal{L}(q,\theta)$ and $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}})$ are the likelihood and KL divergence and are given by

$$\begin{aligned} \mathcal{L}(q,\theta) &= \int_{\mathcal{Y}} q(\mathcal{Y}) \ln\left(\frac{p(\mathcal{X},\mathcal{Y}\,|\,\theta)}{q(\mathcal{Y})}\right) \\ \mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) &= -\int_{\mathcal{Y}} q(\mathcal{Y}) \ln\left(\frac{p(\mathcal{Y}\,|\,\mathcal{X},\theta)}{q(\mathcal{Y})}\right). \end{aligned} \tag{15}$$

It therefore follows (why?) that $\mathcal{L}(q,\theta) \leq l(\theta;\mathcal{X})$ for all distributions, $q(\cdot)$.

Can now use the decomposition of (14) to define the EM algorithm, beginning with an initial parameter estimate, θ_{old}.

The EM Algorithm Revisited

E-Step: Maximize the lower bound, $\mathcal{L}(q, \theta_{old})$, with respect to $q(\cdot)$ while keeping θ_{old} fixed.

In principle this is a variational problem since we are optimizing a functional, but the solution is easily found.

First note that $l(\theta_{old}; \mathcal{X})$ does not depend on $q(\cdot)$.

Then follows from (14) with $\theta = \theta_{old}$ that maximizing $\mathcal{L}(q, \theta_{old})$ is equivalent to minimizing $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}})$.

Since this latter term is always non-negative we see that $\mathcal{L}(q, \theta_{old})$ is optimized when $\mathsf{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) = 0$

– by earlier observation, this is the case when we take $q(\mathcal{Y}) = p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$.

At this point we see that the lower bound, $\mathcal{L}(q, \theta_{old})$, now equals current value of log-likelihood, $l(\theta_{old}; \mathcal{X})$.

The EM Algorithm Revisited

M-Step: Keep $q(\mathcal{Y})$ fixed and maximize $\mathcal{L}(q, \theta)$ over θ to obtain θ_{new}.

This will therefore cause the lower bound to increase (if it is not already at a maximum)

– which in turn means the log-likelihood must also increase.

Moreover, at this new value θ_{new} it will no longer be the case that $\text{KL}(q \,||\, p_{\mathcal{Y}|\mathcal{X}}) = 0$

– so by (14) the increase in the log-likelihood will be greater than the increase in the lower bound.

Comparing Classical EM With General EM

It is instructive to compare the E-step and M-step of the general EM algorithm with the corresponding steps of the original EM algorithm.

To do this, first substitute $q(\mathcal{Y}) = p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$ into (15) to obtain

$$\mathcal{L}(q, \theta) \;=\; Q(\theta; \theta_{old}) \;+\; \text{constant} \tag{16}$$

where $Q(\theta; \theta_{old})$ is the expected complete-date log-likelihood as defined in (1).

The M-step of the general EM algorithm is therefore identical to the M-step of original algorithm since the constant term in (16) does not depend on θ.

The E-step in general EM algorithm takes $q(\mathcal{Y}) = p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$ which, at first glance, appears to be different to original E-step.

But there is no practical difference: original E-step simply uses $p(\mathcal{Y} \,|\, \mathcal{X}, \theta_{old})$ to compute $Q(\theta; \theta_{old})$ and, while not explicitly stated, the general E-step must also do this since it is required for the M-step.

E.G. Imputing Missing Data (Again)

N respondents were asked to answer m questions each. The observed data are:

$$v_{iq} = \begin{cases} 1 & \text{if respondent } i \text{ answered yes to question } q \\ 0 & \text{if respondent } i \text{ answered no to question } q \\ - & \text{if respondent } i \text{ did not answer question } q \end{cases}$$

$$y_{iq} = \begin{cases} 1 & v_{iq} \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

We assume the following model:

- K classes of respondents: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ with $\pi_k = \mathsf{P}(\text{respondent in class } k)$
- Latent variables $z_i \in \{1, \ldots, K\}$ for $i = 1, \ldots, N$
- Class dependent probability of answers: $\sigma_{kq} = \mathbb{P}(v_{iq} = 1 \mid z_i = k)$
- Parameters $\theta = (\boldsymbol{\pi}, \boldsymbol{\sigma})$

Log-likelihood with $\mathcal{X} := \{v_{iq} \,|\, i = 1, \ldots, N, q = 1, \ldots, m\}$:

$$l(\theta; \mathcal{X}) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \prod_{q: y_{iq}=1} \sigma_{kq}^{v_{iq}} (1 - \sigma_{kq})^{(1 - v_{iq})} \right)$$

Question: What implicit assumptions are we making here?

EM for Imputing Missing Data

Take $\mathcal{Y} := (z_1, \ldots, z_N)$.

Complete-data log-likelihood then given by

$$l(\theta; \mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{N} \sum_{k=1}^{K} 1_{\{z_i = k\}} \ln \left(\pi_k \prod_{q: y_{iq}=1} \sigma_{kq}^{v_{iq}} (1 - \sigma_{kq})^{(1-v_{iq})} \right)$$

E-Step: Need to compute $Q(\theta; \theta_{old})$. We have

$$\begin{aligned} Q(\theta; \theta_{old}) &= \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old} \right] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}^{old} \ln \left(\pi_k \prod_{q: y_{iq}=1} \sigma_{kq}^{v_{iq}} (1 - \sigma_{kq})^{(1-v_{iq})} \right) \end{aligned}$$

where

$$\begin{aligned} \gamma_{ik}^{old} := \mathsf{P}(z_i = k \mid \mathbf{v}_i, \theta_{old}) &\propto \pi_k^{old} \mathsf{P}(\mathbf{v}_i \mid z_i = k) \\ &= \pi_k^{old} \prod_{q: y_{iq}=1} (\sigma_{kq}^{old})^{v_{iq}} (1 - \sigma_{kq}^{old})^{(1-v_{iq})} \end{aligned}$$

EM for Imputing Missing Data

M-Step: Now solve for $\theta_{new} = \max_\theta \; Q(\theta; \theta_{old})$:

We have

$$
\begin{aligned}
Q(\theta;\theta_{old}) \quad &= \quad \sum_{k=1}^{K} \left(\sum_{i=1}^{N} \gamma_{ik}^{old} \right) \ln(\pi_k) + \sum_{k=1}^{K} \sum_{q=1}^{m} \left(\sum_{i:y_{iq}=1} \gamma_{ik}^{old} v_{iq} \right) \ln\left(\sigma_{kq}\right) \\
&\quad + \left(\sum_{i:y_{iq}=1} \gamma_{ik}^{old} (1 - v_{iq}) \right) \ln\left(1 - \sigma_{kq}\right)
\end{aligned}
$$

Solving $\max_\theta \; Q(\theta; \theta_{old})$ yields

$$
\begin{aligned}
\pi_k^{new} \quad &= \quad \frac{\sum_{i=1}^{N} \gamma_{ik}^{old}}{\sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij}^{old}} \\
\sigma_{kq}^{new} \quad &= \quad \frac{\sum_{i:y_{iq}=1} \gamma_{ik}^{old} v_{iq}}{\sum_{i:y_{iq}=1} \gamma_{ik}^{old}}.
\end{aligned}
$$

Now iterate E- and M-steps until convergence.