ML estimation in exponential family and the EM algorithm

In exponential family, the underlying pdf or pmf is

$$f(x|\underline{\theta}) = c(\underline{\theta}) \cdot e^{\sum_{j=1}^{k} \theta_j t_j(x)} \cdot h(x)$$

where $\underline{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$ is *canonical parameter*. Then, based on the i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_n)$, the *canonical sufficient statistic* is

$$t(\mathbf{X}) = (\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)) := (t_1(\mathbf{X}), \ldots, t_k(\mathbf{X})),$$

which is also complete (if Θ contains k-dimensional parallelepiped), and therefore it is a minimal sufficient statistic.

Proposition 1 *Under the usual regularity conditions, in exponential families the likelihood equation boils down to solving*

$$\mathbb{E}_{\underline{\theta}}(t(\mathbf{X})) = t(\mathbf{X}).$$

Proof. The likelihood-function has the following form:

$$L_\theta(\mathbf{X}) = c^n(\underline{\theta}) \cdot e^{\sum_{j=1}^{k} \theta_j \sum_{i=1}^{n} t_j(X_i)} \cdot \prod_{i=1}^{n} h(x_i) = \frac{1}{a(\underline{\theta})} \cdot e^{\underline{\theta} \cdot t^T(\mathbf{X})} \cdot b(\mathbf{X}),$$

where the vectors are rows, ^{T} denotes the transposition, and

$$a(\underline{\theta}) = \int_{\mathcal{X}} e^{\underline{\theta} \cdot t^T(\mathbf{x})} \cdot b(\mathbf{x})\, d\mathbf{x}. \tag{1}$$

is the normalizing constant, while $\mathcal{X} \subset \mathbb{R}^n$ is the sample space. This formula will play a crucial role in our subsequent calculations.

The likelihood equation is

$$\nabla_{\underline{\theta}} \ln L_{\underline{\theta}}(\mathbf{X}) = \mathbf{0},$$

that is

$$-\nabla_{\underline{\theta}} \ln a(\underline{\theta}) + \nabla_{\underline{\theta}}(t(\mathbf{X})\underline{\theta}^T) = \mathbf{0}. \tag{2}$$

Under certain regularity conditions, by (1) we get that

$$\nabla_{\underline{\theta}} \ln a(\underline{\theta}) = \frac{1}{a(\underline{\theta})} \int_{\mathcal{X}} t(\mathbf{x}) e^{t(\mathbf{x})\underline{\theta}^T} \cdot b(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{\underline{\theta}}(t(\mathbf{X})).$$

Therefore, (2) is equivalent to

$$-\mathbb{E}_{\underline{\theta}}(t(\mathbf{X})) + t(\mathbf{X}) = \mathbf{0}$$

that finishes the proof. □

Note that this resembles the idea of the moment estimation. Indeed, if $t_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$, …, $t_k(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, then the ML-estimator of the canonical parameter is the same as the moment-estimator. This is the case, e.g., when our underlying distribution is Poisson, exponential, or Gaussian.

EM-ALGORITHM

For the detailed description see [3]. To solve the likelihood equation is sometimes tedious, especially when there are missing data. Instead of numerical methods, the following iteration works:

1. **E**-step: based on the actual value of the parameter, we reconstruct the missing data via taking conditional expectation;

2. **M**-step: from the so completed data we maximize the likelihood in $\underline{\theta}$. With this new $\underline{\theta}$ we go back to step **E**.

Under general conditions, e.g., in exponential families, the iteration converges. Sometimes not the data themselves are missing, but some model parameters (e.g., membership vectors when we want to decompose mixtures).

Notation: let $\mathcal{X}$ and $\mathcal{Y}$ denote the complete and incomplete sample spaces, between which the

$$\mathcal{X} \to \mathcal{Y}, \qquad \mathbf{x} \to \mathbf{y}(\mathbf{x})$$

known many-one mapping works. Denoting by $f(\mathbf{x}|\underline{\theta})$ and $g(\mathbf{y}|\underline{\theta})$ the joint density of the complete and incomplete sample, respectively,

$$g(\mathbf{y}|\underline{\theta}) = \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x}|\underline{\theta})\, d\mathbf{x} \tag{3}$$

where $\mathcal{X}(\mathbf{y}) = \{\mathbf{x} : \mathbf{y}(\mathbf{x}) = \mathbf{y}\} \subset \mathcal{X}$. We want to maximize $g(\mathbf{y}|\underline{\theta})$ in $\underline{\theta}$, based on the incomplete observation $\mathbf{y}$.

Example

In Rao [6] (Section 5.5.g.), the phenotype of 197 offsprings can be AB, Ab, aB, or ab with respective probabilities $\frac{1}{2}+\frac{1}{4}\pi$, $\frac{1}{4}-\frac{1}{4}\pi$, $\frac{1}{4}-\frac{1}{4}\pi$, and $\frac{1}{4}\pi$, where π is the unknown parameter to be estimated based on the counts:

$$\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34).$$

$\mathbf{y}$ follows multinomial distribution with mass function (incomplete likelihood):

$$g(\mathbf{y}|\pi) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{1}{4}\pi)^{y_1}(\frac{1}{4}-\frac{1}{4}\pi)^{y_2}(\frac{1}{4}-\frac{1}{4}\pi)^{y_3}(\frac{1}{4}\pi)^{y_4}.$$

Maximizing g in π is numerically not tractable, therefore, for technical purposes, we complete our data into an $\mathbf{x}$, by splitting y_1 into two parts:

$$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5), \quad \text{where} \quad y_1 = x_1 + x_2, \quad y_2 = x_3, \quad y_3 = x_4, \quad y_4 = x_5.$$

$\mathbf{x}$ also follows multinomial distribution with mass function (complete likelihood):

$$f(\mathbf{x}|\pi) = \frac{(x_1 + x_2 + x_3 + x_4 + x_5)!}{x_1! x_2! x_3! x_4! x_5!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} p_5^{x_5},$$

where

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{4}\pi, \quad p_3 = p_4 = \frac{1}{4} - \frac{1}{4}\pi, \quad p_5 = \frac{1}{4}\pi.$$

Instead of integration, in this discrete situation, we have summation:

$$g(\mathbf{y}|\pi) = \sum_{x_1+x_2=y_1,\, x_1\geq 0,\, x_2\geq 0 \text{ integer},\, x_3=y_2,\, x_4=y_3,\, x_5=y_4} f(\mathbf{x}|\pi).$$

Starting with $\pi^{(0)}$, the iteration is as follows. If we already have $\pi^{(m)}$, the $(m+1)$-th step of the iteration:

1. **E**-step: given $\mathbf{y}$, we find $\mathbf{x}$, namely x_1 and x_2. Given $y_1 = 125$, the conditional distribution of x_1 and x_2 is $\mathcal{B}in_{125}\left(\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}\pi^{(m)}}\right)$ and $\mathcal{B}in_{125}\left(\frac{\frac{1}{4}\pi^{(m)}}{\frac{1}{2}+\frac{1}{4}\pi^{(m)}}\right)$. The conditional expectations are therefore the binomial expectations:

$$x_1^{(m)} = 125 \cdot \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}\pi^{(m)}} \quad \text{és} \quad x_2^{(m)} = 125 \cdot \frac{\frac{1}{4}\pi^{(m)}}{\frac{1}{2} + \frac{1}{4}\pi^{(m)}}.$$

2. **M**-step: based on the the complete data $(x_1^{(m)}, x_2^{(m)}, 18, 20, 34)$, we maximize f in π.

$$f(\mathbf{x}|\pi) = \text{constant} \cdot \left(\frac{1}{4}\pi\right)^{x_2^{(m)}+34} \cdot \left(\frac{1}{4} - \frac{1}{4}\pi\right)^{18+20}.$$

Multiplying with $4^{x_2^{(m)}+34+18+20}$, we have to maximize

$$\tilde{f}(\mathbf{x}|\pi) = \text{const} \cdot (\pi)^{x_2^{(m)}+34} \cdot (1-\pi)^{18+20}$$

in π. Since it resembles the binomial likelihood, the solution is

$$\pi^{(m+1)} = \frac{x_2^{(m)} + 34}{x_2^{(m)} + 34 + 18 + 20}$$

that will be the new value of π.

With this new $\pi^{(m+1)}$, we go back to the **E**-step. Starting with $\pi^{(0)} = 0.5$, after 2-3 steps π stabilized around 0.6.

Theoretical considerations

In exponential family, the complete data likelihood is:

$$f(\mathbf{x}|\underline{\theta}) = c^n(\underline{\theta}) \cdot e^{\sum_{j=1}^{k} \theta_j \sum_{i=1}^{n} t_j(x_i)} \cdot \prod_{i=1}^{n} h(x_i) = \frac{1}{a(\underline{\theta})} \cdot e^{\underline{\theta} \cdot t^T(\mathbf{x})} \cdot b(\mathbf{x}).$$

The E-M iteration works with the sufficient statistic t, in view of Proposition 1. Since the observable incomplete observations $\mathbf{Y}$ are functions of the unobservable complete ones $\mathbf{X}$, the conditional density of $\mathbf{X}$ at $\mathbf{x}$, conditioned on $\mathbf{Y} = \mathbf{y}$, in view of (3) is

$$k(\mathbf{x}|\mathbf{y}, \underline{\theta}) = \frac{f(\mathbf{x}|\underline{\theta})}{g(\mathbf{y}|\underline{\theta})} = \frac{1}{a(\underline{\theta}|\mathbf{y})} \cdot e^{\underline{\theta} \cdot t^T(\mathbf{x})} \cdot b(\mathbf{x}), \tag{4}$$

where

$$a(\underline{\theta}|\mathbf{y}) = \int_{\mathcal{X}(\mathbf{y})} e^{\underline{\theta} \cdot t^T(\mathbf{x})} \cdot b(\mathbf{x})\, d\mathbf{x}. \tag{5}$$

Therefore, the unconditioned and conditioned likelihoods can be written in terms of the same canonical sufficient statistic and parameter, with the exception, that their domains are $\mathcal{X}$ and $\mathcal{X}(\mathbf{y})$, as you see it from (1) and (5).

We want to maximize the log-likelihood function $L(\underline{\theta}) := \ln g(\mathbf{y}|\underline{\theta})$ in $\underline{\theta}$, given $\mathbf{y}$. In view of Proposition 1,

$$\nabla_{\underline{\theta}} L(\underline{\theta}) = -\mathbb{E}(t|\underline{\theta}) + \mathbb{E}(t|\mathbf{y}, \underline{\theta}) \tag{6}$$

and we look for its root in $\underline{\theta}$.

Having $\underline{\theta}^{(m)}$, the iteration is:

1. **E**-step: based on $\underline{\theta}^{(m)}$, we estimate the sufficient statistic t of the complete sample from the incomplete one:

$$t^{(m)} := \mathbb{E}(t|\mathbf{y}, \underline{\theta}^{(m)}). \tag{7}$$

2. **M**-step: we find $\underline{\theta}^{(m+1)}$, as the root of the complete likelihood equation:

$$\nabla_{\underline{\theta}} \ln f(\mathbf{x}|\underline{\theta}) = \mathbf{0}.$$

But by Proposition 1, we have to solve

$$\mathbb{E}(t|\underline{\theta}) = t^{(m)}. \tag{8}$$

Its solution is $\underline{\theta}^{(m+1)}$.

If the iteration converges to $\underline{\theta}^*$, for m sufficiently large, we approximately have $\underline{\theta}^{(m)} = \underline{\theta}^{(m+1)} = \underline{\theta}^*$, and so, in view of (7) and (8),

$$\mathbb{E}(t|\underline{\theta}^*) = \mathbb{E}(t|\mathbf{y}, \underline{\theta}^*)$$

holds, i.e., $\underline{\theta}^*$ is the root of (6).

Now, more generally (not only in exponential family) we will show the convergence of the GEM (General EM) iteration: in the **M**-step we not necessarily maximize, but just increase the objective function. Instead of natural log, because of information theoretical considerations, we use log for the logarithm of base 2, and equivalently, we want to maximize $L(\underline{\theta}) = \log g(\mathbf{y}|\underline{\theta})$.

Let us introduce the following notation: for $\underline{\theta}, \underline{\theta}'$

$$Q(\underline{\theta}'|\underline{\theta}) := \mathbb{E}(\log f(\mathbf{x}|\underline{\theta}')|\mathbf{y}, \underline{\theta}) = \int_{\mathcal{X}(\mathbf{y})} \log f(\mathbf{x}|\underline{\theta}') k(\mathbf{x}|\mathbf{y}, \underline{\theta})\, d\mathbf{x}. \tag{9}$$

With this, the $\underline{\theta}^{(m)} \to \underline{\theta}^{(m+1)}$ phase of the iteration is:

1. **E**-step: calculate $Q(\underline{\theta}|\underline{\theta}^{(m)})$ with taking conditional expectation, just follow (9). (In exponential family, it can be done through the canonical sufficient statistic.)

2. **M**-step: maximize $Q(\underline{\theta}|\underline{\theta}^{(m)})$ in $\underline{\theta}$, and define

   $$\underline{\theta}^{(m+1)} := \arg\max Q(\underline{\theta}|\underline{\theta}^{(m)}).$$

   Assume that $\underline{\theta}^{(m+1)} \in \Theta$.

We will show that the following relaxation of the EM algorithm converges. In the **M**-step, instead of maximizing $Q(\underline{\theta}|\underline{\theta}^{(m)})$ in $\underline{\theta}$, we just increase its value compared to the preceding iteration step. That is, $\underline{\theta}^{(m+1)}$ satisfies

$$Q(\underline{\theta}^{(m+1)}|\underline{\theta}^{(m)}) \geq Q(\underline{\theta}^{(m)}|\underline{\theta}^{(m)}). \tag{10}$$

We further introduce the notation

$$H(\underline{\theta}'|\underline{\theta}) := \mathbb{E}(\log k(\mathbf{x}|\mathbf{y}, \underline{\theta}')|\mathbf{y}, \underline{\theta}) = \int_{\mathcal{X}(\mathbf{y})} \log k(\mathbf{x}|\mathbf{y}, \underline{\theta}') k(\mathbf{x}|\mathbf{y}, \underline{\theta})\, d\mathbf{x}. \tag{11}$$

Lemma 1

$$H(\underline{\theta}'|\underline{\theta}) \leq H(\underline{\theta}|\underline{\theta})$$

with equality if and only if $k(\mathbf{x}|\mathbf{y}, \underline{\theta}) = k(\mathbf{x}|\mathbf{y}, \underline{\theta}')$ *almost surely.*

Proof.

$$H(\underline{\theta}|\underline{\theta}) - H(\underline{\theta}'|\underline{\theta}) = \int_{\mathcal{X}(\mathbf{y})} \log \frac{k(\mathbf{x}|\mathbf{y},\underline{\theta})}{k(\mathbf{x}|\mathbf{y},\underline{\theta}')} k(\mathbf{x}|\mathbf{y},\underline{\theta})\, d\mathbf{x}$$

is the relative entropy of the $k(\mathbf{x}|\mathbf{y},\underline{\theta})$ distribution with respect to the $k(\mathbf{x}|\mathbf{y},\underline{\theta}')$ distribution, and by Lesson 1, it is nonnegative (it is 0 if and only if the two distributions are the same almost surely). □

Definition 1 *The $\underline{\theta}^{(m+1)} = M(\underline{\theta}^{(m)})$ iteration defines a GEM algorithm if*

$$Q(M(\underline{\theta})|\underline{\theta}) \geq Q(\underline{\theta}|\underline{\theta}), \qquad \forall \underline{\theta} \in \Theta.$$

Therefore, when (10) holds, we have a GEM algorithm at hand.

Theorem 1 *For any GEM algorithm*

$$L(M(\underline{\theta})) \geq L(\underline{\theta}), \qquad \forall \underline{\theta} \in \Theta,$$

with equality if and only if $k(\mathbf{x}|\mathbf{y}, M(\underline{\theta})) = k(\mathbf{x}|\mathbf{y},\underline{\theta})$ and $Q(M(\underline{\theta})|\underline{\theta}) = Q(\underline{\theta}|\underline{\theta})$ almost surely.

Proof.

$Q(\underline{\theta}|\underline{\theta}') - H(\underline{\theta}|\underline{\theta}') = \mathbb{E}(\log(f(\mathbf{x}|\underline{\theta}) - \log(k(\mathbf{x}|\mathbf{y},\underline{\theta})|\mathbf{y},\underline{\theta}') = \mathbb{E}(\log(g(\mathbf{y}|\underline{\theta}))|\mathbf{y},\underline{\theta}') = \log(g(\mathbf{y}|\underline{\theta})) = L(\underline{\theta})$

as $\log(g(\mathbf{y}|\underline{\theta}))$ is measurable with respect to $\mathbf{y}$. Therefore,

$$L(M(\underline{\theta})) - L(\underline{\theta}) = [Q(M(\underline{\theta})|\underline{\theta}) - Q(\underline{\theta}|\underline{\theta})] + [H(\underline{\theta}|\underline{\theta}) - H(M(\underline{\theta})|\underline{\theta})] \geq 0,$$

since in the first [] we have a nonnegative quantity by the definition of the GEM, and the quantity in the second [] is also nonnegative by Lemma 1. □

If the likelihood-function is bounded, the monotonous increasing sequence of the GEM – will converge. For further conditions, to have local and global maximum, see [3].

I. Csiszár and P. Shields in [2] prove that the EM-algorithm is also an alternating minimizer of the I-divergence or relative entropy, see Lesson 1.

Assume that the unobservable full sample $x_1^n \in A^n$ is from an unknown distribution $\mathbb{Q} \in \mathcal{Q}$ (feasible set of distributions on the finite alphabet A). We only have the unobservable incomplete sample $y_1^n \in B^n$ ($|B| \leq |A|$). There is a known mapping $T : A \to B$ such that $Tx_i = y_i$.

Starting with an arbitrary $\mathbb{Q}_0 \in \mathcal{Q}$, the EM-iteration is as follows:

- **E**-step: Knowing $\mathbb{Q}_{m-1}$, let $\mathbb{P}_m = \mathbb{E}_{\mathbb{Q}_{m-1}}(\hat{\mathbb{P}}_n|y_1^n)$, where $\hat{\mathbb{P}}_n$ is the empirical distribution of the unobserved full sample. We take the conditional expectation, pretending that the true distribution is $\mathbb{Q}_{m-1}$.

- **M**-step: Calculate the MLE of the distribution of the full sample, pretending that the empirical distribution of x_1^n equals P_m. This MLE will be Q_m.

It can be shown that during the iteration,

$$D(\mathbb{P}_1\|\mathbb{Q}_0) \geq D(\mathbb{P}_1\|\mathbb{Q}_1) \geq D(\mathbb{P}_2\|\mathbb{Q}_1) \geq D(\mathbb{P}_2\|\mathbb{Q}_2) \geq \dots$$

where starting from the distribution $\mathbb{Q}_0$, the sequence $\mathbb{Q}_1, \mathbb{Q}_2, \dots$ reconstructs the unknown distribution of the complete sample. Indeed, in the M-step, $D(\mathbb{P}_m\|\mathbb{Q}_{m-1}) \geq D(\mathbb{P}_m\|\mathbb{Q}_m)$, since by Lemma 3.1, the MLE minimizes the $D(\mathbb{P}_m\|\mathbb{Q})$ I-divergence in $\mathbb{Q} \in \mathcal{Q}$ even if $\mathbb{P}_m$ is not a possible empirical distribution of the full sample. (However, the M-step is easy to perform.) In the E-step, $D(\mathbb{P}_{m-1}\|\mathbb{Q}_{m-1}) \geq D(\mathbb{P}_m\|\mathbb{Q}_{m-1})$ is obtained by the lumping property (see Chapter 4), discussed in Section 5.3 of the lecture notes. Since the sequence of I-divergences is positive and decreasing, the iteration converges. If the set $\mathcal{Q}$ of feasible distributions is convex and compact, the iteration will converge to a global optimum, starting with any $\mathbb{Q}_0 \in \mathcal{Q}$ of maximal support. Otherwise, the iteration can get stuck at a local optimum. The method is non-parametric in that here not the parameter but the distribution itself is estimated.

Application of the EM-algorithm for decomposition of mixtures

EM-algorithm for decomposing Gaussian mixtures

We follow the description of Hastie et al. [4]. If the empirical density of our continuous observations is bimodal, we suspect that it is the mixture of two Gaussian distributions. Let Y be the mixture of $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, while Δ is Bernoulli distributed with parameter π: when $\Delta = 0$, then Y_1, and when when $\Delta = 1$, then Y_2 is the true distribution. The mixture model is

$$Y = (1 - \Delta)Y_1 + \Delta Y_2$$

with parameters (μ_j, σ_j^2) $(j = 1, 2)$ and π, collected in

$$\underline{\theta} = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \pi).$$

The pdf of Y is

$$g(y|\underline{\theta}) = (1 - \pi)f_1(y) + \pi f_2(y),$$

where f_j is the pdf of the $\mathcal{N}(\mu_j, \sigma_j^2)$-distribution.

Based on the n-element sample, $y_1, \dots, y_n$, the likelihood function is

$$g(\mathbf{y}|\underline{\theta}) = \prod_{i=1}^{n} g(y_i|\underline{\theta}) = \prod_{i=1}^{n} [(1-\pi)f_1(y_i) + \pi f_2(y_i)]$$

the logarithm of which is complicated to maximize $\underline{\theta}$. Instead, we apply the EM-iteration, where g is the pdf of the incomplete sample (the complete likelihood would be a product if we knew the memberships of the sample entries), see [4].

0. Initialization:

$$\underline{\theta}^{(0)} = (\mu_1^{(0)}, \sigma_1^{2(0)}, \mu_2^{(0)}, \sigma_2^{2(0)}, \pi^{(0)}).$$

For example, $\pi^{(0)}$ can be $1/2$, the two expectations can be two far sample values, and the variances can be both the empirical. $m := 0$ and assume that we have $\underline{\theta}^{(m)} = (\mu_1^{(m)}, \sigma_1^{2(m)}, \mu_2^{(m)}, \sigma_2^{2(m)}, \pi^{(m)})$. The next step of the inner cycle:

1. E-step: for each sample entry we calculate its contribution to the two components: $\mathbb{E}(\Delta \,|\, Y = y_i) = \mathbb{P}(\Delta = 1 \,|\, Y = y_i)$, and denote it with $\pi_i^{(m+1)}$ $(i = 1, \dots, n)$. By the Bayes rule:

$$\pi_i^{(m+1)} = \frac{\pi^{(m)} f_2^{(m)}(y_i)}{(1-\pi^{(m)}) f_1^{(m)}(y_i) + \pi^{(m)} f_2^{(m)}(y_i)} \qquad (i = 1, \dots, n).$$

2. M-step: we maximize the two Gaussian likelihoods separately in the usual way so that we count the sample entries with their contributions to the components:

$$\mu_1^{(m+1)} = \frac{\sum_{i=1}^n (1-\pi_i^{(m+1)}) y_i}{\sum_{i=1}^n (1-\pi_i^{(m+1)})}, \qquad \sigma_1^{2(m+1)} = \frac{\sum_{i=1}^n (1-\pi_i^{(m+1)})(y_i - \mu_1^{(m+1)})^2}{\sum_{i=1}^n (1-\pi_i^{(m+1)})}$$

and

$$\mu_2^{(m+1)} = \frac{\sum_{i=1}^n \pi_i^{(m+1)} y_i}{\sum_{i=1}^n \pi_i^{(m+1)}}, \qquad \sigma_2^{2(m+1)} = \frac{\sum_{i=1}^n \pi_i^{(m+1)} (y_i - \mu_2^{(m+1)})^2}{\sum_{i=1}^n \pi_i^{(m+1)}}.$$

Then

$$\pi^{(m+1)} := \frac{1}{n} \sum_{i=1}^{n} \pi_i^{(m+1)},$$

$m := m + 1$, and start a new outer cycle. With a "not too bad starting" the iteration will converge, and it can be generalized to the decomposition of more than two components.

EM-algorithm for decomposing polynomial (multinomial) mixtures

We follow the description of [5], sometimes called collaborative filtering. The model is for decomposition of contingency tables into k layers according to a latent (missing) variable.

The incomplete (observable) sample space is $X \times Y$, where $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$ and the counts for the x_i, y_j pairs are collected into an $n \times m$ contingency table with general entry $\nu(x_i, y_j) \geq 0$ (they are usually, but not necessarily, integers). For example, microarrays or keyword-document matrices. These are the missing data, and completed with the latent discrete variable taking on values in $Z = \{z_1, \ldots, z_k\}$. For example, k different tissues or topics (k is fixed, and usually much smaller than n or m). Our purpose is to decompose the table into k layers.

The exact model is the following:

$$p(x_i, y_j) = \sum_{l=1}^{k} p(x_i, y_j|z_l) \cdot \pi(z_l) = \sum_{l=1}^{k} p(x_i|z_l) \cdot p(y_j|z_l) \cdot \pi(z_l)$$

where $p(x_i, y_j)$ denotes the probability of the x_i, y_j pair, $\pi(z_l)$ is the probability (proportion) of the component z_l, and we make the following conditional independence assumption:

$$p(x_i, y_j|z_l) = p(x_i|z_l) \cdot p(y_j|z_l).$$

The model parameters are: $\pi(z_l)$ $(l = 1, \ldots, k)$, $(\sum_{l=1}^{k} \pi(z_l) = 1)$; $p(x_i|z_l)$, $p(y_j|z_l)$ $(i = 1 \ldots, n;\ j = 1, \ldots, m;\ l = 1, \ldots, k)$. We collect them in $\underline{\theta}$. Our purpose is to maximize the following incomplete likelihood (mixture of polynomial distributions):

$$\sum_{l=1}^{k} \pi(z_l) \cdot c_l \prod_{i=1}^{n} \prod_{j=1}^{m} p(x_i, y_j|z_l)^{\nu(x_i, y_j|z_l)},$$

where the conditional cell frequencies $\nu(x_i, y_j|z_l)$'s are not integers any more, and the constant c_l (depends only on l) contains factorials or Γ-functions. We estimate the parameters via the following EM-iteration:

0. Initialization: $\pi^{(0)}(z_l)$, $p^{(0)}(x_i|z_l)$, $p^{(0)}(y_j|z_l)$. t:=0, and assume that we already have $\underline{\theta}^{(t)}$.

1. E-step: we find the conditional expectation of z_l's conditioned on the cell observations. Since $z_1, \ldots, z_k$ are k alternatives, we apply the Bayes rule:

$$p^{(t+1)}(z_l|x_i, y_j) = \frac{p^{(t)}(x_i, y_j|z_l) \cdot \pi^{(t)}(z_l)}{\sum_{l'=1}^{k} p^{(t)}(x_i, y_j|z_{l'}) \cdot \pi^{(t)}(z_{l'})} = \frac{p^{(t)}(x_i|z_l) \cdot p^{(t)}(y_j|z_l) \cdot \pi^{(t)}(z_l)}{\sum_{l'=1}^{k} p^{(t)}(x_i|z_{l'}) p^{(t)}(y_j|z_{l'}) \cdot \pi^{(t)}(z_{l'})}$$

$(i = 1, \ldots, n;\ j = 1, \ldots, m)$.
2. M-step: we maximize the parameters of the truncated polynomial distributions for $l = 1, \ldots k$, separately. For this purpose, we maximize

$$c_l \prod_{i=1}^{n} \prod_{j=1}^{m} p(x_i, y_j|z_l)^{\frac{\nu(x_i,y_j)\cdot p^{(t+1)}(z_l|x_i,y_j)}{h_l}}$$

where the constant h_l is the sum of the terms in the numerator (for i, j). By the conditional independence, we have to maximize

$$c_l \left[\prod_{i=1}^{n} \prod_{j=1}^{m} \{p(x_i|z_l) \cdot p(y_j|z_l)\}^{\nu(x_i,y_j)\cdot p^{(t+1)}(z_l|x_i,y_j)} \right]^{\frac{1}{h_l}}$$

in $p(x_i|z_l)$, $p(y_j|z_l)$ for fixed l $(l = 1, \ldots k)$. Rearranging, and using the classical polynomial ML-estimate, we get the new parameter values

$$p^{(t+1)}(x_i|z_l) = \frac{\sum_{j=1}^{m} \nu(x_i, y_j) \cdot p^{(t+1)}(z_l|x_i, y_j)}{\sum_{i'=1}^{n} \sum_{j=1}^{m} \nu(x_{i'}, y_j) \cdot p^{(t+1)}(z_l|x_{i'}, y_j)} \qquad (i = 1, \ldots, n)$$

and

$$p^{(t+1)}(y_j|z_l) = \frac{\sum_{i=1}^{n} \nu(x_i, y_j) \cdot p^{(t+1)}(z_l|x_i, y_j)}{\sum_{i=1}^{n} \sum_{j'=1}^{m} \nu(x_i, y_{j'}) \cdot p^{(t+1)}(z_l|x_i, y_{j'})} \qquad (j = 1, \ldots, m).$$

Then

$$\pi^{(t+1)}(z_l) := \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} p^{(t+1)}(z_l|x_i, y_j)}{nm} \qquad (l = 1, \ldots, k)$$

$t := t + 1$ and go back to the E-step. Depending on the starting parameter value, $\underline{\theta}^{(t)}$ will converge to a local argmax $\underline{\theta}^*$ of the above likelihood function.

EM-algorithm for estimating the parameters of the stochastic block-model for graphs

We follow the description of [1]. The assumptions of the model are the following. Given a simple graph $G = (V, \mathbf{A})$ ($|V| = n$, with adjacency matrix $\mathbf{A}$) and k $(1 < k < n)$, we are looking for the hidden k-partition $(V_1, \ldots, V_k)$ of the vertices such that

- vertices are independently assigned to cluster V_a with probability π_a, $a = 1, \ldots, k$; $\sum_{a=1}^{k} \pi_a = 1$;

- given the cluster memberships, vertices of V_a and V_b are connected independently, with probability

$$\mathbb{P}(i \sim j \,|\, i \in V_a, j \in V_b) = p_{ab}, \quad 1 \le a, b \le k.$$

The parameters are collected in the vector $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ and the $k \times k$ symmetric matrix $\mathbf{P}$ of p_{ab}'s.

Our statistical sample is the $n \times n$ symmetric, 0-1 adjacency matrix $\mathbf{A} = (a_{ij})$ of G. There are no loops, so the diagonal entries are zeros. Based on $\mathbf{A}$, we want to estimate the parameters of the above block model.

Using the theorem of mutually exclusive and exhaustive events, the likelihood function is the mixture of joint distributions of i.i.d. Bernoulli distributed entries:

$$\begin{aligned} &\frac{1}{2} \sum_{1 \le a,b \le k} \pi_a \pi_b \prod_{i \in V_a, j \in V_b, i \ne j} p_{ab}^{a_{ij}} (1 - p_{ab})^{(1 - a_{ij})} \\ &= \frac{1}{2} \sum_{1 \le a,b \le k} \pi_a \pi_b \cdot p_{ab}^{e_{ab}} (1 - p_{ab})^{(n_{ab} - e_{ab})}. \end{aligned}$$

This is the mixture of binomial distributions, where e_{ab} is the number of edges connecting vertices of V_a and V_b ($a \ne b$), while e_{aa} is twice the number of edges with both endpoints in V_a; further,

$$n_{ab} = |V_a| \cdot |V_b| \quad (a \ne b) \quad \text{and} \quad n_{aa} = |V_a| \cdot (|V_a| - 1) \quad (a = 1, \ldots, k) \tag{12}$$

are the numbers of possible edges between V_a, V_b and within V_a, respectively.

Here $\mathbf{A}$ is the incomplete data specification as the cluster memberships are missing. Therefore, it is straightforward to use the EM-algorithm, for parameter estimation from incomplete data.

First we complete our data matrix $\mathbf{A}$ with latent membership vectors $\Delta_1, \ldots, \Delta_n$ of the vertices that are k-dimensional i.i.d. $Poly(1, \underline{\pi})$ (polinomially distributed) random vectors. More precisely, $\Delta_i = (\Delta_{1i}, \ldots, \Delta_{ki})$, where $\Delta_{ai} = 1$ if $i \in V_a$ and zero otherwise. Thus, the sum of the coordinates of any Δ_i is 1, and $\mathbb{P}(\Delta_{ai} = 1) = \pi_a$.

Based on these, the likelihood function above is

$$\frac{1}{2} \sum_{1 \le a,b \le k} \pi_a \pi_b \cdot p_{ab}^{\sum_{i \ne j} \Delta_{ai} \Delta_{bj} a_{ij}} \cdot (1 - p_{ab})^{\sum_{i \ne j} \Delta_{ai} \Delta_{bj} (1 - a_{ij})}$$

that is maximized in the alternating **E** and **M** steps of the EM-algorithm.

Note that that the complete likelihood would be the squareroot of

$$\begin{aligned}&\prod_{1\le a,b\le k} p_{ab}^{e_{ab}}\cdot(1-p_{ab})^{(n_{ab}-e_{ab})}\\&=\prod_{a=1}^{k}\prod_{i=1}^{n}\prod_{b=1}^{k}[p_{ab}^{\sum_{j:\,j\neq i}\Delta_{bj}a_{ij}}\cdot(1-p_{ab})^{\sum_{j:\,j\neq i}\Delta_{bj}(1-a_{ij})}]^{\Delta_{ai}}\end{aligned}\tag{13}$$

that is valid only in case of known cluster memberships.

Starting with initial parameter values $\underline{\pi}^{(0)}$, $\mathbf{P}^{(0)}$ and membership vectors $\Delta_1^{(0)},\dots,\Delta_n^{(0)}$, the t-th step of the iteration is the following ($t=1,2,\dots$).

- **E**-step: we calculate the conditional expectation of each Δ_i conditioned on the model parameters and on the other cluster assignments obtained in step $t-1$ and collectively denoted by $M^{(t-1)}$. By the Bayes theorem, the responsibility of vertex i for cluster a is

$$\begin{aligned}\pi_{ai}^{(t)}&=\mathbb{E}(\Delta_{ai}\,|\,M^{(t-1)})\\&=\frac{\mathbb{P}(M^{(t-1)}|\Delta_{ai}=1)\cdot\pi_a^{(t-1)}}{\sum_{b=1}^{k}\mathbb{P}(M^{(t-1)}|\Delta_{bi}=1)\cdot\pi_b^{(t-1)}}\end{aligned}$$

  ($a=1,\dots,k$; $i=1,\dots,n$). For each i, $\pi_{ai}^{(t)}$ is proportional to the numerator, where

$$\begin{aligned}&\mathbb{P}(M^{(t-1)}|\Delta_{ai}=1)\\&=\prod_{b=1}^{k}(p_{ab}^{(t-1)})^{\sum_{j\neq i}\Delta_{bj}^{(t-1)}a_{ij}}\cdot(1-p_{ab}^{(t-1)})^{\sum_{j\neq i}\Delta_{bj}^{(t-1)}(1-a_{ij})}\end{aligned}$$

  is the part of the likelihood (13) effecting vertex i under the condition $\Delta_{ai}=1$.

- **M**-step: we maximize the truncated binomial likelihood

$$p_{ab}^{\sum_{i\neq j}\pi_{ai}^{(t)}\pi_{bj}^{(t)}a_{ij}}\cdot(1-p_{ab})^{\sum_{i\neq j}\pi_{ai}^{(t)}\pi_{bj}^{(t)}(1-a_{ij})}$$

  with respect to the parameter p_{ab}, for all a,b pairs separately. Obviously, the maximum is attained by the following estimators of p_{ab}'s comprising the symmetric matrix $\mathbf{P}^{(t)}$: $p_{ab}^{(t)}=\frac{\sum_{i,j:\,i\neq j}\pi_{ai}^{(t)}\pi_{bj}^{(t)}a_{ij}}{\sum_{i,j:\,i\neq j}\pi_{ai}^{(t)}\pi_{bj}^{(t)}}$ ($1\le a\le b\le k$), where edges connecting vertices of clusters a and b are counted fractionally, multiplied by the membership probabilities of their endpoints.

The maximum likelihood estimator of $\underline{\pi}$ in the t-th step is $\underline{\pi}^{(t)}$ of coordinates $\pi_a^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \pi_{ai}^{(t)}$ $(a = 1, \ldots, k)$, while that of the membership vector Δ_i is obtained by discrete maximization: $\Delta_{ai}^{(t)} = 1$ if $\pi_{ai}^{(t)} = \max_{b\in\{1,\ldots,k\}} \pi_{bi}^{(t)}$ and 0, otherwise. (In case of ambiguity, the cluster with the smallest index is selected.) This choice of $\underline{\pi}$ will increase (better to say, not decrease) the likelihood function. Note that it is not necessary to assign vertices uniquely to the clusters, the responsibility π_{ai} of a vertex i can as well be regarded as the intensity of vertex i belonging to cluster a.

According to the general theory of the EM-algorithm, in exponential families (as in the present case), convergence to a local maximum can be guaranteed (depending on the starting values), but it runs in polynomial time in the number of vertices n. However, the speed and limit of the convergence depends on the starting clustering, which can be chosen by means of preliminary application of some nonparametric multiway cut algorithm or spectral clustering methods, see [1].

Bibliography

[1] Bolla, M., Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables. Wiley (2013).

[2] Csiszár, I., Shields, P., Information Theory and Statistics: A Tutorial, In: Foundations and Trends in Communications and Information Theory, Vol. 1 Issue 4 (2004), Now Publishers, USA.

[3] Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. B* 39 (1977) 1-38.

[4] Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, New York (2001).

[5] Hofmann, T., Puzicha, J., Latent class models for collaborative filtering, in Proc. of IJCAI'99 (1999).

[6] Rao, C. R., Linear Statistical Inference and Its Applications, Wiley, New York (1965, 1973).