EM algorithm: an exmaple with mixture probabilistic PCA

Wei Liu

2021-01-15

mixture probabilistic PCA

Consider a model

$$x_i = \mu_k + W_k z_i + \varepsilon_k, \quad if\ y_{ik} = 1, y_{ij} = 0, i \neq j. \tag{1.1}$$

where $x_i \in R^p, z_i \in R^q, \varepsilon_k \sim N(0, \sigma_k^2 I_p), z_i \sim N(0, I_q), W_k \in R^{p\times q}, y_i \sim Multinormial(1, \pi), \pi \in R^K$. First, it performs clustering, then conducts PCA. The following contents are divided into three parts:

1) evaluate full-data loglikelihood $l(\theta)$;
2) take posterior expectation of latent variables on $l(\theta)$, and obtain Q-function;
3) Maximize the Q-function.

Full-data loglikelihood

By (1.1), we can obtain the complete-data likelihood for individual i, $P(x_i, z_i, y_i) = p(x_i|z_i, y_i)p(z_i|y_i)p(y_i)$, whose specific form is

$$\Pi_{k=1}^K \{\pi_k f_{ki} g_i\}^{y_{ik}},$$

where

$$f_{ki} = (2\pi\sigma_k^2)^{-p/2} \exp\{\frac{-1}{2\sigma_k^2}\|x_i - \mu_k - W_k z_i\|^2\}, g_i = (2\pi)^{-q/2} \exp(-\frac{1}{2}\|z_i\|^2).$$

The corresponding loglikelihood is given by

$$l = \sum_{k=1}^K y_{ik}\{\ln(\pi_k) + \ln f_{ki} + \ln g_i\} \tag{1.2}$$

EM algorithm

The essential objective of EM algorithm is to maximize the observation likelihood. To deduce EM algorithm, we first calculate the posterior distribution of (y_i, z_i) given x_i and parameters by previous iteration. Noting each y_{ik} is seperatable, we consider

$$P(y_{ik} = 1, z_i|x_i) = P(y_{ik} = 1|x_i)P(z_i|y_{ik} = 1, x_i) = R_{ik} f_k(z_i),$$

where $f_k(z_i)$ is the posterior distribution of z_i given x_i when $y_{ik} = 1$, $R_{ik} = P(y_{ik} = 1|x_i) = a_{ki}/(\sum_k a_{ki})$, where $a_{ki} = \pi_k |C_k|^{-1/2} \exp\{-\frac{1}{2}(x_i - \mu_k)^T C_k^{-1}(x_i - \mu_k)\}$ and $C_k = \sigma_k^2 I + W_k W_k^T$. By (8) in Bishop (1999), we have

$$f_k(z_i) = (2\pi)^{-q/2} |\sigma_k^2 M_k|^{1/2} \exp\{(z_i - s_k(x_i)^T \sigma_k^{-2} M_k(z_i - s_k(x_i))\},$$

where $\sigma_k^2 M_k^{-1} = \sigma_k^2(\sigma_k^2 I_q + W_k^T W_k)^{-1}$ and $s_k(x_i) = M_k^{-1} W_k(x_i - \mu_k)$.

E-step

We rewrite (1.2) as the specific form,

$$l_k = y_{ik}\{\ln(\pi_k) - \frac{p}{2}\ln(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}\|x_i - \mu_k - W_k z_i\|^2 - \frac{q}{2}\ln(2\pi) - \frac{1}{2}\|z_i\|^2\}.$$

Omitting the terms independent of parameters, we have

$$l_k = y_{ik}\{\ln(\pi_k) - \frac{p}{2}\ln(\sigma_k^2) - \frac{1}{2\sigma_k^2}\|x_i - \mu_k - W_k z_i\|^2 - \frac{1}{2}\|z_i\|^2\}.$$

Thus

$$E_{(y_{ik},z_i)|x_i}(l_k) = \int l_k(y_{ik}, z_i)P(y_{ik} = 1, z_i|x_i)d(y_{ik}, z_i) = \int l_k(1, z_i)R_{ik}f_k(z_i)dz_i. \tag{1.3}$$

Denote $h_k(z_i) = \ln(\pi_k) - \frac{p}{2}\ln(\sigma_k^2) - \frac{1}{2\sigma_k^2}\|x_i - \mu_k - W_k z_i\|^2 - \frac{1}{2}\|z_i\|^2 = \ln(\pi_k) - \frac{p}{2}\ln(\sigma_k^2) - \frac{1}{2}z_i^T z_i - \frac{1}{2\sigma_k^2}\{z_i^T W^T W z_i - 2(x_i - \mu_k)^T W_k z_i + \|x_i - \mu_k\|^2\}$. Furthermore, (1.3) simplifies as

$$E_{(y_{ik},z_i)|x_i}(l_k) = R_{ik}\int h_k(z_i)f_k(z_i)dz_i. \tag{1.4}$$

(1.4) only involves the posterior first-order moment and second-order moment of z_i that are denoted by

$$\langle z_i\rangle = M_k^{-1}W_k^T(x_i - \mu_k)$$

and

$$\langle z_i z_i^T\rangle = \sigma_k^2 M_k^{-1} + \langle z_i\rangle\langle z_i\rangle^T.$$

Similar to (54) in Bishop (1999), we obtain

$$E_{(y_i,z_i)|x_i}l = \sum_{k=1}^{K} E_{(y_{ik},z_i)|x_i}(l_k) = \sum_{k=1}^{K} R_{ik}\{\ln(\pi_k) - \frac{p}{2}\ln(\sigma_k^2) - \frac{1}{2}\langle z_i z_i^T\rangle - \frac{1}{2\sigma_k^2}(tr(W_k^T W_k\langle z_i z_i^T\rangle) - 2(x_i - \mu_k)^T W_k\langle z_i\rangle + \|x_i - \mu_k\|^2)\}$$

Finally, we obtain the Q-function,

$$Q(\theta;\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} R_{ik}(\theta^{(t)})\{\ln(\pi_k) - \frac{p}{2}\ln(\sigma_k^2) - \frac{1}{2}tr(\langle z_i z_i^T\rangle) - \frac{1}{2\sigma_k^2}(tr(W_k^T W_k\langle z_i z_i^T\rangle) - 2(x_i - \mu_k)^T W_k\langle z_i\rangle + \|x_i - \mu_k\|^2)\},$$

where $\langle z_i\rangle$ and $\langle z_i z_i^T\rangle$ also include $\theta^{(t)}$.

M-step

This step is to maximize the Q-function. Denote $\theta = (\pi_k, \sigma_k^2, \mu_k, W_k, k \le K)$, all involved parameters. Since the constraint $\sum_{k=1}^{K}\pi_k = 1$ is required, we use Langrange method to obtain a new objective function,

$$L(\theta, \lambda; \theta^{(t)}) = Q(\theta; \theta^{(t)}) + \lambda(1 - \sum_{k=1}^{K}\pi_k).$$

1) Taking derivative on π_k, λ, and setting it to zero, we obtain

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{n} R_{ik}(\theta^{(t)})\pi_k^{-1} - \lambda = 0 \tag{2.2.1}$$

$$\sum_{k=1}^{K}\pi_k = 1 \tag{2.2.2}$$

Combining (2.2.1) and (2.2.2), we conclude

$$\pi_k^{(t+1)} = n^{-1}\sum_{i=1}^{n} R_{ik}(\theta^{(t)})$$

by using the fact that $\sum\limits_{i=1}^{n}(\sum\limits_{k=1}^{K} R_{ik}(\theta^{(t)})) = n$.

2) Taking derivative on μ_k, we have

$$\mu_k^{(t+1)} = \frac{\sum\limits_{i=1}^{n} R_{ik}(\theta^{(t)})\{x_i - W_k^{(t+1)}\langle z_i\rangle^{(t)}\}}{\sum\limits_{i=1}^{n} R_{ik}(\theta^{(t)})}$$

3) Taking derivative on W_k and using scalar-to-Matrix derivative, we get

$$\sum_{i=1}^{n}[R_{ik}(\theta^{(t)})\{\langle z_i\rangle(x_i-\mu_k)^T - \langle z_i z_i^T\rangle W_k^T\}] = 0$$

which leads to

$$W_k^{(t+1)} = \sum_{i=1}^{n}[R_{ik}(\theta^{(t)})(x_i-\mu_k^{(t+1)})\langle z_i\rangle^T][\sum_{i=1}^{n}[R_{ik}(\theta^{(t)})\langle z_i z_i^T\rangle]^{-1}. \tag{2.2.3}$$

4) Denote $s_{ik}(W_k,\mu_k) = tr(W_k^T W_k\langle z_i z_i^T\rangle) - 2(x_i-\mu_k)^T W_k\langle z_i\rangle + \|x_i-\mu_k\|^2$. Taking derivative on σ_i^2, we get

$$\sigma_k^{2,(t+1)} = \frac{\sum\limits_{i=1}^{n} R_{ik}(\theta^{(t)}) s_{ik}(W_k^{(t+1)},\mu_k^{(t+1)})}{p\sum\limits_{i=1}^{n} R_{ik}(\theta^{(t)})}. \tag{2.2.4}$$

Two-stage EM procedure

Note that M-step equations for μ_i and W_i are coupled, so further manipulation is required to obtain explicit solutions.

The likelihood function we wish to maximize is given by

$$L(\theta) = \sum_{i=1}^{n}\ln\big\{\sum_{k=1}^{K}\pi_k p(x_i|y_{ik}=1)\big\}.$$

Now, we introduce labels y_i as missing data, and ignore the presence of the latent z_i. Here, z_i is integrated, so only y_i is missing data. Then the "full" loglikelihood is

$$L(\theta;x,y) = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\ln\big\{\pi_k p(x_i|y_{ik}=1)\big\}.$$

Based on this full log likelihood, we will construct EM algorithm. Thus, the expected complete-data log likelihood is given by

$$\hat{L} = \sum_{i=1}^{n}\sum_{k=1}^{K} R_{ik}\ln\big\{\pi_k p(x_i|y_{ik}=1)\big\}, \tag{2.3.1}$$

from which we get the updation of $\pi_k^{(t+1)}$ and $\mu_k^{(t+1)}$:

$$\pi_k^{(t+1)} = n^{-1} \sum_{i=1}^{n} R_{ik}(\theta^{(t)})$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} R_{ik}(\theta^{(t)}) x_i}{\sum_{i=1}^{n} R_{ik}(\theta^{(t)})}.$$

However, we don't solve σ_k^2 and W_k from (2.3.1), because there is no closed-form in it. Actually, we only need to find $\sigma_k^{2,(t+1)}$ and $W_k^{(t+1)}$ increasing $\hat{L}(\theta)$. (2.2.3) and (2.2.4) based on $L(\theta, \lambda; \theta^{(t)})$ regarding z_i and y_i as missing data provide the iterative value such that condition.

We update W_k by

$$W_k^{(t+1)} = \sum_{i=1}^{n} [R_{ik}(\theta^{(t)})(x_i - \mu_k^{(t+1)}) \langle z_i \rangle^T] [\sum_{i=1}^{n} [R_{ik}(\theta^{(t)}) \langle z_i z_i^T \rangle]^{-1}.$$

and update σ_k^2 by

$$\sigma_k^{2,(t+1)} = \frac{\sum_{i=1}^{n} R_{ik}(\theta^{(t)}) s_{ik}(W_k^{(t+1)}, \mu_k^{(t+1)})}{p \sum_{i=1}^{n} R_{ik}(\theta^{(t)})}.$$

So far, each parameter has a iterative closed-form solution.

Convergence check

Since EM algorithm is a subclass of MM algorithm, by the principle of MM algorithm we can check the convergence by the fact that

$$Q(\theta; \theta^{(t)}) \leq L(\theta) \quad forall \quad \theta$$

and

$$Q(\theta; \theta^{(t)}) = L(\theta) \; if \quad and \; only \; if \; \theta = \theta^{(t)}.$$

Thus, we have

$$L(\theta^{(t)}) = Q(\theta^{(t)}; \theta^{(t)}) \leq Q(\theta^{(t+1)}; \theta^{(t)}) \leq L(\theta^{(t+1)}). \tag{2.4.1}$$

Recursively, we have

$$Q(\theta^{(t+1)}; \theta^{(t)}) \leq Q(\theta^{(t+2)}; \theta^{(t+1)}). \tag{2.4.2}$$

Thus, there are two methods to check the convergence (correction of programming) of algorithm from the aspect of the objective function.

1) By (2.4.1), we can check whether the value of the observed loglikelihood function is nondecreasing.
2) By (2.4.2), we can check whether the value of the Q-function is nondecreasing.

Generalized EM algorithm

We learn the generalized EM algorithm in this section, whose definition is referred to Dempster (1977, JRSSB, EM and GEM). An iterative algorithm $\theta^{(t+1)} = M(\theta^{(t)})$ is a generalized EM if

$$Q(M(\theta); \theta) \geq Q(\theta; \theta).$$

So, we only need that θ iterates one step towards the nondecreasing direction of $Q(\theta;\theta^{(t)})$. MM algorithm is an extension of GEM in the sense that Q-function is changed to 1)Minorization function, i.e. $L(\theta) \geq Q(\theta;\theta^{(t)})$ and 2)equality holds iif $\theta = \theta^{(t)}$.

Assume $\{y_i, i \leq n\}$ is the observed data, $\{z_i, i \leq n\}$ the latent variable, and we are interested in parameter θ.

Following the principle of EM algorithm, the complete-data log likelihood is given by

$$l(\theta; Y, Z) = \sum_i \ln(P(y_i, z_i; \theta)).$$

Next, according the posterior distribution of z_i given y_i, $P(z_i|y_i;\theta)$, we take conditional expectation on z_i for $l(\theta; Y, Z)$ to obtain Q function. However, it is often difficult to calculate $P(z_i|y_i;\theta)$ in practice, which leads to that EM algorithm fails. In this backgroud, GEM is developed to solve this problem.

First, we inspect the another derivation of EM algorithm,

$$\ln P(Y;\theta) = \ln P(Y, Z;\theta) - \ln P(Z|Y;\theta) = \ln \frac{P(Y,Z;\theta)}{q(Z)} - \ln \frac{P(Z|Y;\theta)}{q(Z)}, \tag{3.1}$$

where $q(Z)$ is the density function of Z and is a unknown function to be optimized. Taking expectation with respect to Z on both sides of (3.1), we have

$$\ln P(Y;\theta) = \sum_z q(z) \ln \frac{P(Y,z;\theta)}{q(z)} - \sum_z q(z) \ln \frac{P(z|Y;\theta)}{q(z)},$$

where the first term is called evidence lower bound (ELBO), and the second term is KL divergence of $P(Z|Y,\theta)$ and $q(Z)$. That is

$$ELBO = \sum_z q(z) \ln \frac{P(Y,z;\theta)}{q(z)}$$

and

$$KL(q(Z)\|P(Z|Y,\theta)) = \sum_z q(z) \ln \frac{P(z|Y;\theta)}{q(z)}.$$

Thus, we obtain

$$\ln P(Y;\theta) = ELBO + KL(q(Z)\|P(Z|Y,\theta)). \tag{3.2}$$

Recalling EM algorithm, paramter θ is fixed at E-step, so $\ln P(Y;\theta)$ is constant here. Thus, the optimized solution of $q(z)$ is equal to $P(z|y;\theta)$ as much as possible. By this way, the E-step of GEM turns to

$$\arg\max_{q(z)} ELBO$$

due to

$$\arg\min_{q(z)} KL(q(z)\|P(z|Y=y,\theta)) \Leftrightarrow \arg\max_{q(z)} ELBO$$

by the fact that $KL(q(Z)\|P(Z|Y,\theta)) = \ln P(Y;\theta) - ELBO$.

And the M-step of GEM is

$$\theta = \arg\max_\theta ELBO(\theta).$$

In summary, GEM algorithm is given by

$$E-step: \quad q(z)^{(t+1)} = \arg\max_{q(z)} \sum_z q(z) \ln \frac{P(Y,z;\theta^{(t)})}{q(z)}; \tag{3.3}$$

$$M-step: \quad \theta^{(t+1)} = \arg\max_\theta \sum_z q(z)^{(t+1)} \ln \frac{P(Y,z;\theta^{(t)})}{q(z)^{(t+1)}}. \tag{3.4}$$

Given inital value $\theta^{(0)}$, then repeat (3.3) and (3.4) until convergence. Actually, GEM algorithm belongs to the class of coordinate ascent algortihm, that is, EBLO is a bivariant function on $q(z)$ and θ; first, we optimize $q(z)$ given θ; then we optimize θ given $q(z)$.

```
Remark 1: q(z) is also a parameter joining in iteration.
Remark 2: GEM does not involve computing P(z|y;\theta).
Remark 3: In practice, we assume a parametric form for q(z) to approximate P(z|Y;\theta),
          then optimize the parameter in iteration, which is called
          variational Bayesian EM algorithm.
```

See https://mbernste.github.io/posts/elbo/ for more details about GEM and ELBO. Why ELBO is called evidence lower bound? Since, given θ, $\ln P(Y;\theta)$ is called evidence, which indicates the evidence of model fitting data by taking θ. By Jensen inequality, we have $\ln P(Y;\theta) \geq ELBO$, a lower bound of envidence, so ELBO is called envidence lower bound.


```
References:
  https://mbernste.github.io/posts/elbo/
  https://zhuanlan.zhihu.com/p/150342963
  Shi X, Jiao Y, Yang Y, Cheng CY, Yang C, Lin X, Liu J. VIMCO: variational inference for multiple corre
association studies. Bioinformatics. 2019 Oct 1;35(19):3693-3700. doi: 10.1093/bioinformatics/btz167. PM
  Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. Neura
```