Latent Variable Models and EM algorithm

SC4/SM4 Data Mining and Machine Learning, Hilary Term 2017

Dino Sejdinovic

3.1 Clustering and Mixture Modelling

K-means and hierarchical clustering are non-probabilistic algorithms — based on the intuitive notions of clustering "similar" instances together and "dissimilar" instances apart. Their goal is not to model the probability of the observed data items. In contrast, *probabilistic unsupervised learning* constructs a *generative model* that describes clustering of the items. We assume that there is some latent / unobserved process that is governing the data generation - and based on the data, we will try to answer the questions about this generating process.

Mixture models assume that our dataset $\mathbf{X}$ was created by sampling iid from K distinct populations (called *mixture components*). In other words, data come from a mixture of several sources and the model for the data can be viewed as a convex combination of several distinct probability distributions, often modelled with a given parametric family.

Samples in population k can be modelled using a distribution F_{μ_k} with density $f(x|\mu_k)$, where μ_k is the *model parameter* for the k-th component. For a concrete example, consider a p-dimensional multivariate normal density with unknown mean μ_k and *known diagonal* covariance $\sigma^2 I$,

$$f(x|\mu_k) = |2\pi\sigma^2|^{-\frac{p}{2}} \exp\left(-\frac{1}{2\sigma^2}\|x-\mu_k\|_2^2\right). \tag{3.1}$$

Such model corresponds to the following generative model, whereby for each data item $i = 1, 2, \ldots, n$, we

(i) first determine the assignment variable (independently for each data item i):

$$Z_i \overset{i.i.d.}{\sim} \text{Discrete}(\pi_1, \ldots, \pi_K) \qquad \text{i.e., } \mathbb{P}(Z_i = k) = \pi_k$$

where for $k = 1, \ldots, K$, $\pi_k \geq 0$, such that $\sum_{k=1}^K \pi_k = 1$, are the *mixing proportions*, additional model parameters to be inferred;

(ii) then, given the assignment $Z_i = k$ of the mixture component, $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})^\top$ is sampled (independently) from the corresponding k-th component:

$$X_i|(Z_i = k) \sim f(x|\mu_k).$$

We observe $X_i = x_i$ for each i but do not observe its assignment Z_i (*latent variables*), and would like to infer the parameters $\theta = (\mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K)$ as well as the latent variables.

Note that the complete log-likelihood in the model

$$\log p(\mathbf{z}, \mathbf{X}|\theta) = \log\left(\prod_{i=1}^{n} \pi_{z_i} f(x_i|\mu_{z_i})\right) = \sum_{i=1}^{n} \left(\log \pi_{z_i} + \log f(x_i|\mu_{z_i})\right) \tag{3.2}$$

is not available as z_i is not observed. We can consider marginalising over the latent variables

$$p(\mathbf{X}|\theta) = \sum_{z_1=1}^{K} \cdots \sum_{z_n=1}^{K} \prod_{i=1}^{n} \pi_{z_i} f(x_i|\mu_{z_i}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f(x_i|\mu_k)\right). \tag{3.3}$$

giving the *marginal log-likelihood* of the observations,

$$\ell(\theta) = \log p(\mathbf{X}|\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f(x_i|\mu_k).$$

However, direct maximisation is not feasible and the marginal log-likelihood will often have many local optima. Fortunately, there is a simple local marginal log-likelihood maximisation algorithm called Expectation Maximisation (EM), which we will describe in Section 3.3.

3.2 KL Divergence and Gibbs' Inequality

Before we describe the EM algorithm, we will review the notion of *Kullback-Leibler (KL) divergence* or *relative entropy* between probability distributions P and Q.

KL divergence.

- Let P and Q be two absolutely continuous probability distributions on $\mathcal{X} \subseteq \mathbb{R}^d$ with densities p and q respectively. Then the KL divergence *from Q to P* is defined as

$$D_{KL}(P \parallel Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx. \tag{3.4}$$

- Let P and Q be two discrete probability distributions with probability mass functions p and q respectively. Then the KL divergence *from Q to P* is defined as

$$D_{KL}(P \parallel Q) = \sum_{i} p(x_i) \log \frac{p(x_i)}{q(x_i)}. \tag{3.5}$$

In both cases, we can write

$$D_{KL}(P \parallel Q) = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right], \tag{3.6}$$

where $\mathbb{E}_p$ denotes that expectation is taken over p. By convexity of $f(x) = -\log(x)$ and Jensen's inequality (3.8), we have that

$$D_{KL}(P \parallel Q) = \mathbb{E}_p\left[-\log \frac{q(X)}{p(X)}\right] \geq -\log \mathbb{E}_p \frac{q(X)}{p(X)} = 0, \tag{3.7}$$

where in the last step we used that $\int_{\mathcal{X}} q(x)dx = 1$ in continuous case and $\sum_i q(x_i) = 1$ in discrete case.

Jensen's inequality. Let f be a convex function and X be a random variable. Then

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}X\right). \tag{3.8}$$

If f is strictly convex, then equality holds if and only if X is almost surely a constant.

Thus, we conclude that KL-divergence is always non-negative. This consequence of Jensen's inequality is called *Gibbs' inequality*. Moreover, since $f(x) = -\log(x)$ is strictly convex on $x > 0$, the equality holds if and only if $p(x) = q(x)$ almost everywhere, i.e. $P = Q$. Note that in general KL-divergence is *not symmetric*: $D_{KL}\left(P \parallel Q\right) \neq D_{KL}\left(Q \parallel P\right)$.

3.3 EM Algorithm

EM algorithm is a general purpose iterative strategy for local maximisation of the likelihood under missing data/hidden variables. The method has been proposed many times for specific models– it was given its name and studied as a general framework by [1].

Let $(\mathbf{X}, \mathbf{z})$ be a pair of observed variables $\mathbf{X}$, and latent variables $\mathbf{z}$. Our probabilistic model is given by $p(\mathbf{X}, \mathbf{z}|\theta)$, but we have no access to $\mathbf{z}$. Therefore, we would like to maximise the observed data log-likelihood (marginal log-likelihood) $\ell(\theta) = \log p(\mathbf{X}|\theta) = \log \int p(\mathbf{X}, \mathbf{z}|\theta)d\mathbf{z}$ over θ. However, marginalisation of latent variables typically results in an intractable optimization problem and we need to resort to approximations.

Now, assume for a moment that we have access to another objective function $\mathcal{F}(\theta, q)$, where $q(\mathbf{z})$ is a certain distribution on latent variables $\mathbf{z}$, which we are free to choose and will call *variational distribution*. Moreover, assume that $\mathcal{F}$ satisfies

$$\mathcal{F}(\theta, q) \leq \ell(\theta) \text{ for all } \theta,\, q, \tag{3.9}$$

$$\max_q \mathcal{F}(\theta, q) = \ell(\theta), \tag{3.10}$$

i.e. $\mathcal{F}(\theta, q)$ is a *lower bound on the log-likelihood* for any variational distribution q (3.9), which also *matches the log-likelihood* at a particular choice of q (3.10).

Given these two properites, we can construct an alternating maximisation: *coordinate ascent* algorithm as follows:

Coordinate ascent on the lower bound. For $t = 1, 2 \ldots$ until convergence:

$$q^{(t)} := \operatorname*{argmax}_q \mathcal{F}(\theta^{(t-1)}, q)$$

$$\theta^{(t)} := \operatorname*{argmax}_\theta \mathcal{F}(\theta, q^{(t)}).$$

Theorem 3.1. Assuming (3.9) and (3.10), coordinate ascent on the lower bound $\mathcal{F}(\theta, q)$ does not decrease the log likelihood $\ell(\theta)$.

Proof. $\ell(\theta^{(t-1)}) = \mathcal{F}(\theta^{(t-1)}, q^{(t)}) \leq \mathcal{F}(\theta^{(t)}, q^{(t)}) \leq \mathcal{F}(\theta^{(t)}, q^{(t+1)}) = \ell(\theta^{(t)})$. □

Additional assumption, that $\nabla^2_\theta \mathcal{F}(\theta^{(t)}, q^{(t)})$ are negative definite with eigenvalues $< -\epsilon < 0$, implies that $\theta^{(t)} \to \theta^*$ where θ^* is a local MLE.

But how to find such lower bound $\mathcal{F}$? It is given by the so called *variational free energy*, which we define next.

Definition 3.1. *Variational free energy* in a latent variable model $p(\mathbf{X}, \mathbf{z}|\theta)$ is defined as

$$\mathcal{F}(\theta, q) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{z}|\theta) - \log q(\mathbf{z})], \tag{3.11}$$

where q is any probability density/mass function over the latent variables $\mathbf{z}$.

Consider the KL divergence between $q(\mathbf{z})$ and the true conditional based on our model $p(\mathbf{z}|\mathbf{X}, \theta) = p(\mathbf{X}, \mathbf{z}|\theta)/p(\mathbf{X}|\theta)$ for the observations $\mathbf{X}$ and a fixed parameter vector θ. Since KL is non-negative,

$$\begin{aligned} 0 \leq D_{KL}\left[q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{X}, \theta)\right] &= \mathbb{E}_{\mathbf{z}\sim q} \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{X}, \theta)} \\ &= \log p(\mathbf{X}|\theta) + \mathbb{E}_{\mathbf{z}\sim q} \log \frac{q(\mathbf{z})}{p(\mathbf{X}, \mathbf{z}|\theta)}. \end{aligned}$$

Thus, we have obtained a lower bound on the marginal log-likelihood which holds true for *any parameter value* θ and *any choice of the variational distribution* q:

$$\ell(\theta) = \log p(\mathbf{X}|\theta) \geq \mathbb{E}_{\mathbf{z}\sim q} \log \frac{p(\mathbf{X}, \mathbf{z}|\theta)}{q(\mathbf{z})} = \underbrace{\mathbb{E}_{\mathbf{z}\sim q} \log p(\mathbf{X}, \mathbf{z}|\theta)}_{\text{energy}} \overbrace{-\mathbb{E}_{\mathbf{z}\sim q} \log q(\mathbf{z})}^{\text{entropy}}. \tag{3.12}$$

The right hand side in (3.12) is precisely the variational free energy - we see it decomposes in two terms. The first term is usually referred to as *energy* using the physics terminology, more precisely it is the *expected complete data log-likelihood* (if we observed $\mathbf{z}$, we would just maximise the complete data log-likelihood $\log p(\mathbf{X}, \mathbf{z}|\theta)$, but since $\mathbf{z}$ is not observed we need to integrate it out - but recall that q here is *any* distribution over latent variables). The second term is the Shannon entropy $H(q) = -\mathbb{E}_q \log q(\mathbf{z})$ of the variational distribution $q(\mathbf{z})$, and does not depend on θ (it can be thought of as the complexity penalty on q).

The inequality becomes an equality when KL divergence is zero, i.e. when $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{X}, \theta)$ which means that the optimal choice of variational distribution q for fixed parameter value θ is *the true conditional of the latent variables given the observations and that* θ.

Thus, we have proved the following lemma:

Lemma 3.1. Let $\mathcal{F}$ be the variational free energy in a latent variable model $p(\mathbf{X}, \mathbf{z}|\theta)$. Then

- $\mathcal{F}(\theta, q) \leq \ell(\theta)$ for all q and for all θ, and
- For any θ, $\mathcal{F}(\theta, q) = \ell(\theta)$ iff $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \theta)$.

Thus, properties (3.9) and (3.10) are satisfied and we can recast the alternating maximisation of the variational free energy into iterative updates of q (E-step, via the plug-in full conditional of $\mathbf{z}$ using the current estimate of θ) and the updates of θ (M-step, by maximising the 'energy' for the current estimate of q). Provided that both E-step and M-step can be solved exactly, EM Algorithm converges to the local maximum likelihood solution.

EM Algorithm. Initialize $\theta^{(0)}$. At time $t \geq 1$:

- E-step: Set $q^{(t)}(\mathbf{z}) = p(\mathbf{z}|\mathbf{X}, \theta^{(t-1)})$
- M-step: Set $\theta^{(t)} = \arg\max_\theta \mathbb{E}_{\mathbf{z}\sim q^{(t)}} \log p(\mathbf{X}, \mathbf{z}|\theta)$.

3.4 EM Algorithm for Mixtures

Consider again our mixture model from Section 3.1 with

$$p(\mathbf{z}, \mathbf{X}|\theta) = \prod_{i=1}^{n} \pi_{z_i} f(x_i|\mu_{z_i}).$$

Recall that our latent variables $\mathbf{z}$ are discrete (they correspond to cluster assignments) so q is a probability mass function over $\mathbf{z} := (z_i)_{i=1}^n$. Using the expression (3.2), we can write the variational free energy as

$$\begin{aligned}
\mathcal{F}(\theta, q) =& \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{z}|\theta) - \log q(\mathbf{z})] \\
=& \mathbb{E}_q \left[\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}(z_i = k) \left(\log \pi_k + \log f(x_i|\mu_k) \right) \right) - \log q(\mathbf{z}) \right] \\
=& \sum_{\mathbf{z}} q(\mathbf{z}) \left[\left(\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}(z_i = k) \left(\log \pi_k + \log f(x_i|\mu_k) \right) \right) - \log q(\mathbf{z}) \right] \\
=& \sum_{i=1}^{n} \sum_{k=1}^{K} q(z_i = k) \left(\log \pi_k + \log f(x_i|\mu_k) \right) + H(q).
\end{aligned}$$

We will denote $Q_{ik} = q(z_i = k)$, which is called *responsibility of cluster k for data item i.*

Now, the E-step simplifies because

$$p(\mathbf{z}|\mathbf{X},\theta) = \frac{p(\mathbf{X},\mathbf{z}|\theta)}{p(\mathbf{X}|\theta)} = \frac{\prod_{i=1}^n \pi_{z_i} f(x_i|\mu_{z_i})}{\sum_{\mathbf{z}'}\prod_{i=1}^n \pi_{z_i'} f(x_i|\mu_{z_i'})} = \prod_{i=1}^n \frac{\pi_{z_i} f(x_i|\mu_{z_i})}{\sum_k \pi_k f(x_i|\mu_k)}$$
$$= \prod_{i=1}^n p(z_i|x_i,\theta).$$

Thus, for a fixed $\theta^{(t-1)} = (\mu_1^{(t-1)},\dots,\mu_K^{(t-1)},\pi_1^{(t-1)},\dots,\pi_K^{(t-1)})$ we can set

$$Q_{ik}^{(t)} = p(z_i = k|x_i,\theta^{(t-1)}) = \frac{\pi_k^{(t-1)} f(x_i|\mu_k^{(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} f(x_i|\mu_j^{(t-1)})}. \tag{3.13}$$

Now, consider the M-step. For mixing proportions we have a constraint that $\sum_{j=1}^K \pi_j = 1$, so we introduce the Lagrange multiplier and obtain

$$\nabla_{\pi_k}\left(\mathcal{F}(\theta,q) - \lambda(\textstyle\sum_{j=1}^K \pi_j - 1)\right)$$
$$= \sum_{i=1}^n \frac{Q_{ik}}{\pi_k} - \lambda = 0 \quad \Rightarrow \quad \pi_k \propto \sum_{i=1}^n Q_{ik}.$$

Since

$$\sum_{k=1}^K\sum_{i=1}^n Q_{ik} = \sum_{i=1}^n \underbrace{\sum_{k=1}^K Q_{ik}}_{=1} = n,$$

the M-step update for mixing proportions is

$$\pi_k^{(t)} = \frac{\sum_{i=1}^n Q_{ik}^{(t)}}{n}, \tag{3.14}$$

i.e., they are simply given by the total responsibility of each cluster. Note that this update holds regardless of the form of the parametric family $f(\cdot|\mu_k)$ used for mixture components.

Setting derivative with respect to μ_k to 0, we obtain

$$\nabla_{\mu_k}\mathcal{F}(\theta,q) = \sum_{i=1}^n Q_{ik}\nabla_{\mu_k}\log f(x_i|\mu_k) = 0. \tag{3.15}$$

This equation can be solved quite easily for mixture of normals in (3.1), giving the M-step update

$$\mu_k^{(t)} = \frac{\sum_{i=1}^n Q_{ik}^{(t)} x_i}{\sum_{i=1}^n Q_{ik}^{(t)}}, \tag{3.16}$$

which implies that the k-th cluster mean estimate is simply a weighted average of all the data items, where the weights correspond to the responsibilities of cluster k for these points.

Put together, the EM for normal mixture model with known (fixed) covariance is very similar to K-means algorithm where cluster assignments are soft, i.e. rather than assigning each data item x_i to a single cluster at each iteration, we carry forward a responsibility vector $(Q_{i1}, \ldots, Q_{iK})$ giving probabilities of x_i belonging to each cluster. Indeed, K-means algorithm can be undestood as EM where $\sigma^2 \to 0$, such that E-step will assign exactly one entry in $(Q_{i1}, \ldots, Q_{iK})$ to one (corresponding to the nearest mean vector) and the rest to zero.

EM for Normal Mixtures (known covariance) – "Soft K-means"

1. Initialize K cluster means $\mu_1, \ldots, \mu_K$ and mixing proportions $\pi_1, \ldots, \pi_K$.
2. *Update responsibilites (E-step):* For each $i = 1, \ldots, n$, $k = 1, \ldots, K$:
$$Q_{ik} = \frac{\pi_k \exp\left(-\frac{1}{2\sigma^2}\|x_i - \mu_k\|_2^2\right)}{\sum_{j=1}^K \pi_j \exp\left(-\frac{1}{2\sigma^2}\|x_i - \mu_j\|_2^2\right)} \tag{3.17}$$
3. *Update parameters (M-step):* Set $\mu_1, \ldots, \mu_K$ and $\pi_1, \ldots, \pi_K$ and based on the new cluster responsibilities:
$$\pi_k = \frac{\sum_{i=1}^n Q_{ik}}{n}, \qquad \mu_k = \frac{\sum_{i=1}^n Q_{ik} x_i}{\sum_{i=1}^n Q_{ik}}. \tag{3.18}$$
4. Repeat steps 2-3 until convergence.
5. Return the responsibilites $\{Q_{ik}\}$ and parameters $\mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K$.

In some cases, depending on the form of the parametric family $f(\cdot|\mu_k)$ the M-step update for mixtures cannot be solved exactly. In these cases, we can use *gradient ascent* algorithm *inside the M-step*:

$$\mu_k^{(r+1)} = \mu_k^{(r)} + \alpha \sum_{i=1}^n Q_{ik} \nabla_{\mu_k} \log f(x_i|\mu_k^{(r)}).$$

This leads to *generalized EM algorithm*.

3.5 Probabilistic PCA

So far, we have considered the application of EM to clustering, but it can be applied to latent variable models more broadly. Here, we will derive EM for *Probabilistic PCA* [3], a latent variable model for probabilistic dimensionality reduction. Just like in PCA, we try to model a collection of n p-dimensional vectors using a k-dimensional representation with $k < p$. Probabilistic PCA corresponds to the following generative model.

For each data item $i = 1, 2, \ldots, n$:

- Let Y_i be a (latent) k-dimensional normally distributed random vector with mean 0 and identity covariance:
$$Y_i \sim \mathcal{N}(0, I_k),$$
- Given Y_i, the distribution of the i-th data item is a p-dimensional normal:
$$X_i \sim \mathcal{N}(\mu + LY_i, \sigma^2 I)$$
where the parameters $\theta = (\mu, L, \sigma^2)$ correspond to a vector $\mu \in \mathbb{R}^p$, a matrix $L \in \mathbb{R}^{p \times k}$ and $\sigma^2 > 0$.

Note that unlike in clustering, the latent variables $Y_1, \ldots, Y_n$ are now continuous.

From an equivalent representation $X_i = \mu + LY_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_p)$ and is independent of Y, we see that the marginal model on X_i's is

$$f(x|\theta) \;=\; \mathcal{N}\left(x; \mu, LL^\top + \sigma^2 I\right),$$

where parameters are denoted $\theta = \left(\mu, L, \sigma^2\right)$. From here it is clear that the maximum marginal likelihood estimator of μ is available directly as $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ and thus, we do not require EM to estimate μ. We will henceforth assume that the data is centred, to simplify notation and remove μ from the parameters.

On the other hand, maximum marginal likelihood solution for L is unique only up to orthonormal transformations, which is why a certain form of L is usually enforced (e.g. lower-triangular, orthogonal columns). [3] shows that the MLE for PPCA has the following form. Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of the sample covariance and $V_{1:k} \in \mathbb{R}^{p \times k}$ the top k eigenvectors as before. Let $Q \in \mathbb{R}^{k \times k}$ be any orthogonal matrix. Then we have:

$$\mu^{\mathrm{MLE}} = \bar{x} \qquad (\sigma^2)^{\mathrm{MLE}} = \tfrac{1}{p-k}\textstyle\sum_{j=k+1}^p \lambda_j$$
$$L^{\mathrm{MLE}} = V_{1:k}\,\mathrm{diag}((\lambda_1 - (\sigma^2)^{\mathrm{MLE}})^{\frac{1}{2}}, \ldots, (\lambda_k - (\sigma^2)^{\mathrm{MLE}})^{\frac{1}{2}})Q.$$

We note that the standard PCA is recovered when $\sigma^2 \to 0$. However, the EM algorithm we derive below can be faster than eigendecomposition, can be implemented online, can handle missing data and can be extended to more complicated models. We will now proceed by deriving the EM algorithm.

E-step. By Gaussian conditioning (*exercise*),

$$q(y_i) = p(y_i|x_i, \theta) = \mathcal{N}\left(y_i|b_i, R\right),$$

where

$$b_i \;=\; \left(L^\top L + \sigma^2 I\right)^{-1} L^\top x_i, \tag{3.19}$$
$$R \;=\; \sigma^2 \left(L^\top L + \sigma^2 I\right)^{-1}. \tag{3.20}$$

M-step. Recall that the parameters of interest are $\theta = \left(L, \sigma^2\right)$ (since the marginal maximum likelihood estimate of the mean parameter μ is directly available). We would like to maximise the variational free energy given by:

$$\mathcal{F}(\theta, q) \quad = \quad \mathbb{E}_{\mathbf{y}\sim q}\left[\sum_{i=1}^{n} \log p(x_i, y_i|\theta)\right] + \text{const.}$$

By ignoring terms that do not depend on θ and denoting $=_c$ to mean "equal up to a constant independent on θ"

$$\begin{aligned}
\log p(x_i, y_i|\theta) &=_c -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(x_i - Ly_i\right)^\top \left(x_i - Ly_i\right) \\
&=_c -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\{x_i^\top x_i - 2x_i^\top Ly_i + y_i^\top L^\top Ly_i\right\} \\
&=_c -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\{x_i^\top x_i - 2x_i^\top Ly_i + \text{Tr}\left[L^\top Ly_i y_i^\top\right]\right\}.
\end{aligned}$$

Taking expectation over $q(y_i) = \mathcal{N}(y_i|b_i, R)$ gives

$$\mathbb{E}_{y_i\sim q}\left(\log p(x_i, y_i|\theta)\right) =_c -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\{x_i^\top x_i - 2x_i^\top Lb_i + \text{Tr}\left[L^\top L\left(b_i b_i^\top + R\right)\right]\right\}.$$

It remains to sum over all observations to get

$$\begin{aligned}
\mathcal{F}(\theta, q) \quad =_c \quad & -\frac{np}{2}\log\sigma^2 \\
& -\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n} x_i^\top x_i - 2\sum_{i=1}^{n} x_i^\top Lb_i + \text{Tr}\left[L^\top L\left(\sum_{i=1}^{n} b_i b_i^\top + nR\right)\right]\right\}.
\end{aligned}$$

Now, we have

$$\frac{\partial\mathcal{F}}{\partial L} \quad = \quad \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} x_i b_i^\top - L\left(\sum_{i=1}^{n} b_i b_i^\top + nR\right)\right\},$$

which by setting to 0 gives the update rule

$$L^{(\text{new})} = \left(\sum_{i=1}^{n} x_i b_i^\top\right)\left(\sum_{i=1}^{n} b_i b_i^\top + nR\right)^{-1}. \tag{3.21}$$

Letting $\tau = \sigma^{-2}$, we have:

$$\frac{\partial\mathcal{F}}{\partial\tau} \quad = \quad \frac{np}{2}\frac{1}{\tau} - \frac{1}{2}\left\{\sum_{i=1}^{n} x_i^\top x_i - 2\sum_{i=1}^{n} x_i^\top Lb_i + \text{Tr}\left[L^\top L\left(\sum_{i=1}^{n} b_i b_i^\top + nR\right)\right]\right\},$$

and thus

$$\left(\sigma^2\right)^{(\text{new})} = \frac{1}{np}\left\{\sum_{i=1}^{n} x_i^\top x_i - 2\sum_{i=1}^{n} x_i^\top L^{(new)} b_i + \text{Tr}\left[L^{(\text{new})\top} L^{(\text{new})}\left(\sum_{i=1}^{n} b_i b_i^\top + nR\right)\right]\right\}. \tag{3.22}$$

Both Probabilistic PCA and normal mixtures are examples of linear Gaussian models, all of which have the corresponding learning algorithms based on EM. For a unifying review of these and a number of other models from the same family, including factor analysis and hidden Markov models, cf. [2].

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[2] Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural Comput.*, 11(2):305–345, February 1999.

[3] M. E. Tipping and Christopher Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21/3:611–622, January 1999.