T.61.5140 Machine Learning: Advanced Probablistic Methods
Hollmén, Raiko (Spring 2008)
Problem session, 25th of April, 2008
http://www.cis.hut.fi/Opinnot/T-61.5140/

1. Show that the EM algorithm is a special case of the VB-EM algorithm where the family of approximate distributionis $q(\boldsymbol{\theta})$ for the parameters is restricted to delta distributions (distributions where the whole probability mass is concentrated on a single point). Some assumptions have to be made: The family of approximate distribution $q(\mathbf{Z})$ for latent variables $\mathbf{Z}$ should include the true posterior $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$. KL-divergence will go to infinity, so the minimization has to be considered as a limiting process (in practice by ignoring the term $q \ln q$ that can be considered constant). Also, VB-EM usually has a prior for $\boldsymbol{\theta}$, while EM does not. Let us consider the version of EM with a prior for the parameters.

EM algorithm with a prior for parameters:

$$q(\mathbf{Z}) \leftarrow p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \tag{1}$$
$$\boldsymbol{\theta} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} E_{q(\mathbf{Z})} \{\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\} \tag{2}$$

VB-EM algorithm:

$$q(\mathbf{Z}) \leftarrow \operatorname*{argmin}_{q(\mathbf{Z})} E_{q(\boldsymbol{\theta})} \{\mathrm{KL}\left(q(\mathbf{Z}) \parallel p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})\right)\} \tag{3}$$
$$q(\boldsymbol{\theta}) \leftarrow \operatorname*{argmin}_{q(\boldsymbol{\theta})} E_{q(\mathbf{Z})} \{\mathrm{KL}\left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z})\right)\} \tag{4}$$

Kullback-Leibler divergence:

$$\mathrm{KL}\left(q(x) \parallel p(x)\right) = E_{q(x)} \left\{ \ln \frac{q(x)}{p(x)} \right\} \tag{5}$$

Solution:

Starting from the VB-E step and assuming that $q(\boldsymbol{\theta})$ is a delta distribu-

tion we get:

$$q(\mathbf{Z}) \leftarrow \underset{q(\mathbf{Z})}{\operatorname{argmin}} E_{q(\boldsymbol{\theta})} \left\{ \mathrm{KL} \left(q(\mathbf{Z}) \parallel p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \right) \right\} \tag{6}$$

$$= \underset{q(\mathbf{Z})}{\operatorname{argmin}} \mathrm{KL} \left(q(\mathbf{Z}) \parallel p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}) \right) \tag{7}$$

$$= p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}), \tag{8}$$

where the first step is the expectation over the delta distribution, which is just the value at the single point where the mass is concentrated. The second step follows from the fact that KL-divergence is minimized (becomes 0) when the two distributions are the same.

Starting from the VB-M step and assuming that $q(\boldsymbol{\theta})$ is a delta distribution we get:

$$q(\boldsymbol{\theta}) \leftarrow \underset{q(\boldsymbol{\theta})}{\operatorname{argmin}} E_{q(\mathbf{Z})} \left\{ \mathrm{KL} \left(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z}) \right) \right\} \tag{9}$$

$$= \underset{q(\boldsymbol{\theta})}{\operatorname{argmin}} E_{q(\mathbf{Z}, \boldsymbol{\theta})} \left\{ \ln q(\boldsymbol{\theta}) - \ln p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z}) \right\} \tag{10}$$

$$= \underset{q(\boldsymbol{\theta})}{\operatorname{argmin}} E_{q(\mathbf{Z}, \boldsymbol{\theta})} \left\{ - \ln p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Z}) \right\} \tag{11}$$

$$= \underset{q(\boldsymbol{\theta})}{\operatorname{argmax}} E_{q(\mathbf{Z}, \boldsymbol{\theta})} \left\{ \ln \frac{p(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Z})}{p(\mathbf{X}, \mathbf{Z})} \right\} \tag{12}$$

$$= \underset{q(\boldsymbol{\theta})}{\operatorname{argmax}} E_{q(\mathbf{Z}, \boldsymbol{\theta})} \left\{ \ln p(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Z}) \right\} \tag{13}$$

$$\boldsymbol{\theta} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmax}} E_{q(\mathbf{Z})} \left\{ \ln p(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Z}) \right\} \tag{14}$$

In the third step we drop the term $E_{q(\boldsymbol{\theta})}\{\ln q(\boldsymbol{\theta})\}$. If we consider distributions q that are getting peakier and peakier at different points $\boldsymbol{\theta}$, this term depends only on the shape of the distribution and not the point $\boldsymbol{\theta}$ where it is located. In that sense it is a constant that can be dropped. (In the limit where it becomes a delta distribution this term goes to infinity.) In the fift step, $p(\mathbf{X}, \mathbf{Z})$ is also dropped because it is a constant with respect to $q(\boldsymbol{\theta})$. The last step is about the delta distribution for $q(\boldsymbol{\theta})$, where we just want to find the single point $\boldsymbol{\theta}$ and the expectation over $q(\boldsymbol{\theta})$ is just the value at the single point.

2. Consider two extensions of probabilistic principal component analysis (PPCA) with mixture models. The model equation $\mathbf{x}_j = \mathbf{A}\mathbf{s}_j + \boldsymbol{\epsilon}_j$ and the noise model $p(\boldsymbol{\epsilon}_j) = N(\boldsymbol{\epsilon}_j \mid \mathbf{0}, v\mathbf{I})$. The parameters are fixed to $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$, $v = 0.01$, and mixture coefficients $\pi = 0.5$ in all cases. The sources $\mathbf{s}_j$ are distributed according to the mixture of Gaussians. The two cases are: (a) The mixing coefficients π_k are shared among the two sources s_{1j} and s_{2j}

$$p(\mathbf{s}_j) = \sum_{k=1}^{2} \pi_k N(\mathbf{s}_j \mid \boldsymbol{\mu}_k, \Sigma_k), \tag{15}$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \tag{16}$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{17}$$

(b) The mixture is done individually to the two sources s_{1j} and s_{2j}.

$$p(s_{1j}) = \sum_{k=1}^{2} \pi_{1k} N(s_{1j} \mid \mu_{1k}, \sigma_{1k}^2), \tag{18}$$

$$p(s_{2j}) = \sum_{k=1}^{2} \pi_{2k} N(s_{2j} \mid \mu_{2k}, \sigma_{2k}^2), \tag{19}$$

$$\mu_{11} = 0, \quad \mu_{12} = 0, \quad \mu_{21} = 0, \quad \mu_{22} = 0 \tag{20}$$

$$\sigma_{11} = 1, \quad \sigma_{12} = 0.3, \quad \sigma_{21} = 1, \quad \sigma_{22} = 0.3. \tag{21}$$

Sketch $p(\mathbf{x}_j \mid \mathbf{A}, v)$ in both cases.

Solution:

First we should note that the noise variance v is so small that we practically need to consider only the prior distribution of the sources and map it to the data space. The mapping is such that source vector $\mathbf{s} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ maps to the data vector $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and the direction of the other other source axis we get from $\mathbf{s} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ that maps to $\mathbf{x} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

The densities are plotted in Figure 1. Case (a) is a clustering model, whereas (b) resembles independent component analysis (ICA).

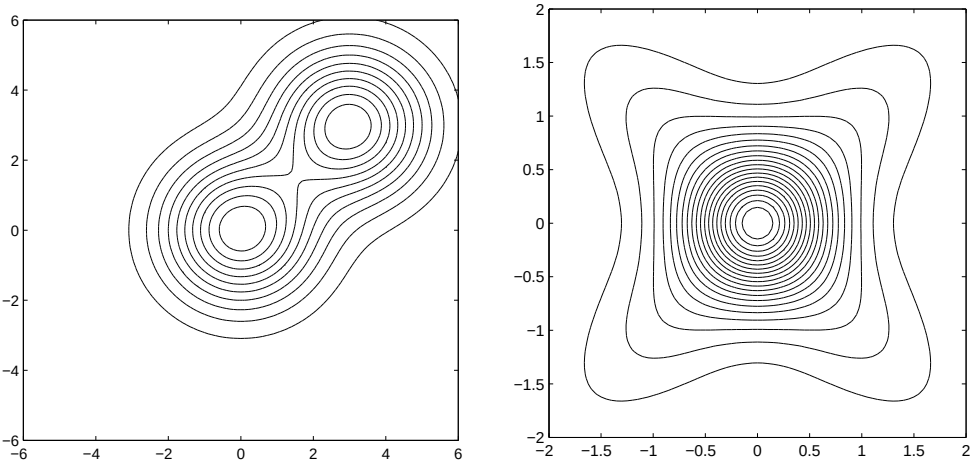

Figure 1: Left: Problem 2(a). Contours of $p(\mathbf{x}_j \mid \mathbf{A}, v)$. Right: Problem 2(b). Contours of $p(\mathbf{x}_j \mid \mathbf{A}, v)$. The horizontal axis is x_{1j} and the vertical axis is x_{2j} in both plots.