CSCI 1950-F: EM for Probabilistic PCA

Brown University, Spring 2011

Supplement to Lecture from April 21, 2011

In this question, we examine the probabilistic principal component analysis (PPCA) model introduced in lecture. To simplify things, we make the following assumptions:

- The N training vectors $x_i \in \mathbb{R}^{D\times 1}$, $i = 1, \ldots, N$, have already been centered, so that $\sum_{i=1}^{N} x_i = 0$. We thus constrain the PPCA model to also have zero mean.
- The desired latent space is one-dimensional, so that observations are represented by coordinates $z_i \in \mathbb{R}$.

In this case, the PPCA generative model can be written as follows:

$$p(z_i) = \text{Normal}(z_i \mid 0, 1) \qquad p(x_i \mid z_i, w, \lambda) = \text{Normal}(x_i \mid wz_i, \lambda I_D)$$

Here, I_D is a $D \times D$ identity matrix, and $w \in \mathbb{R}^{D\times 1}$ and $\lambda > 0$ are parameters to be estimated from the N training observations $x = [x_1, x_2, \ldots, x_N]$. We will estimate these parameters via the EM algorithm.

a) *Suppose that the PPCA model parameters w, λ are known, and consider the posterior distribution $p(z_i \mid x_i, w, \lambda)$. Computation of this distribution is the E-step of the EM algorithm for PPCA. What standard family is it a member of? Give explicit formulas for all parameters of the posterior distribution, in terms of x_i, w, and λ.*

Because $p(z_i)$ is normal, and x_i is a linear function of z_i plus independent Gaussian noise, the posterior $p(z_i \mid x_i)$ is also normal. The two formula sheet expressions for normal conditional distributions provide two alternative forms of this posterior:

$$\begin{aligned} p(z_i \mid x_i, w, \lambda) &= \text{Normal}(z_i \mid (w^T w + \lambda)^{-1} w^T x_i, \lambda/(w^T w + \lambda)) \\ &= \text{Normal}(z_i \mid w^T (ww^T + \lambda I_D)^{-1} x_i, 1 - w^T (ww^T + \lambda I_D)^{-1} w) \end{aligned}$$

The first form is computationally preferable, as it avoids matrix inversion.

b) *Give an expression for the expected complete-data log-likelihood $E[\log p(x, z \mid w, \lambda)]$, where the expectation is with respect to a distribution on z in the family determined in part (a). What particular expectations of z_i must be computed to explicitly evaluate this expression?*

$$\begin{aligned} E[\log p(x, z \mid w, \lambda)] &= \sum_{i=1}^{N} -\frac{1}{2} E[z_i^2] - \frac{D}{2}\log(\lambda) - \frac{1}{2\lambda} E[||x_i - wz_i||^2] \\ &= \sum_{i=1}^{N} -\frac{1}{2}(s_i + m_i^2) - \frac{D}{2}\log(\lambda) - \frac{1}{2\lambda}(||x_i - wm_i||^2 + ||w||^2 s_i) \end{aligned}$$

Here, we have dropped the $1/\sqrt{2\pi}$ normalization constants, and expressed the expected complete-data log-likelihood in terms of $m_i = E[z_i]$ and $s_i = \text{Var}[z_i]$, or equivalently $E[z_i^2] = s_i + m_i^2$. These statistics are directly available from the E-step.

c) *Take the derivative of the expected log-likelihood from part (b) with respect to λ, set to zero, and simplify to determine the M-step estimate $\hat{\lambda}$ of the variance parameter.*

$$\frac{\partial E[\log p(x, z \mid w, \lambda)]}{\partial \lambda} = -\frac{ND}{2\lambda} + \frac{1}{2\lambda^2} \sum_{i=1}^{N} ||x_i - wm_i||^2 + ||w||^2 s_i = 0$$

$$\hat{\lambda} = \frac{1}{ND} \sum_{i=1}^{N} ||x_i - wm_i||^2 + ||w||^2 s_i$$

d) *Take the derivative of the expected log-likelihood from part (b) with respect to w_k, an element of the principal subspace vector w. Set this expression to zero, and simplify to determine the M-step estimate $\hat{w}$ of the principal subspace.*

$$\frac{\partial E[\log p(x, z \mid w, \lambda)]}{\partial w_k} = -\frac{1}{2\lambda} \sum_{i=1}^{N} -2m_i(x_{ik} - w_k m_i) + 2w_k s_i = 0$$

$$\sum_{i=1}^{N} w_k m_i^2 + w_k s_i = \sum_{i=1}^{N} m_i x_{ik}$$

$$\hat{w}_k = \left[\sum_{i=1}^{N} m_i^2 + s_i \right]^{-1} \cdot \sum_{i=1}^{N} m_i x_{ik}$$

e) *Suppose that the EM algorithm converges to a particular set of parameters $\hat{w}, \hat{\lambda}$. Are these ML parameter estimates unique? If so, provide an argument for why this is the case. If not, construct an alternative set of parameters $\bar{w}, \bar{\lambda}$ which have equal log-likelihood, i.e. which satisfy $\log p(x \mid \hat{w}, \hat{\lambda}) = \log p(x \mid \bar{w}, \bar{\lambda})$.*

If we set $\bar{w} = -\hat{w}$ and $\bar{\lambda} = \hat{\lambda}$, we recover an equivalent model because $E[x_i x_i^T] = \hat{w}\hat{w}^T + \hat{\lambda} I_D = \bar{w}\bar{w}^T + \bar{\lambda} I_D$. This change, which effectively reflects the latent space $z_i \to -z_i$, is a special case of the rotational ambiguity seen in higher dimensions.

Useful formulae

$$\int_a^b x^n \, dx = \frac{b^{n+1} - a^{n+1}}{n+1}$$

$$\begin{aligned}
\text{Beta}(\theta \mid a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, \theta^{a-1}\,(1-\theta)^{b-1} \qquad \text{for } 0 \le \theta \le 1 \\
\text{E}[\,\theta \mid a, b\,] &= \int_0^1 \theta\ \text{Beta}(\theta \mid a, b)\, d\theta = \frac{a}{a+b} \\
\text{Var}[\,\theta \mid a, b\,] &= \text{E}[\,\theta^2 \mid a, b\,] - \text{E}[\,\theta \mid a, b\,]^2 = \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

$$\begin{aligned}
\text{Unif}(\theta \mid a, b) &= \begin{cases} 1/(b-a) & \text{if } a \le \theta \le b \\ 0 & \text{otherwise} \end{cases} \\
\text{E}[\,\theta \mid a, b\,] &= \int_{-\infty}^{\infty} \theta\ \text{Unif}(\theta \mid a, b)\, d\theta = \frac{a+b}{2} \\
\text{Var}[\,\theta \mid a, b\,] &= \text{E}[\,\theta^2 \mid a, b\,] - \text{E}[\,\theta \mid a, b\,]^2 = \frac{(b-a)^2}{12}
\end{aligned}$$

$$\begin{aligned}
\text{Normal}(\theta \mid \mu, \lambda) &= \frac{1}{\sqrt{2\pi\lambda}}\ \exp\left\{-\frac{(\theta-\mu)^2}{2\lambda}\right\} \qquad \text{for } \theta \in \mathbb{R} \\
\text{E}[\,\theta \mid \mu, \lambda\,] &= \int_{-\infty}^{\infty} \theta\ \text{Normal}(\theta \mid \mu, \lambda)\, d\theta = \mu \\
\text{Var}[\,\theta \mid \mu, \lambda\,] &= \text{E}[\,\theta^2 \mid \mu, \lambda\,] - \text{E}[\,\theta \mid \mu, \lambda\,]^2 = \lambda
\end{aligned}$$

$$\begin{aligned}
\text{Normal}(\theta \mid \mu, \Lambda) &= \frac{1}{(2\pi)^{d/2}|\Lambda|^{1/2}}\ \exp\left\{-\frac{1}{2}(\theta-\mu)^T\Lambda^{-1}(\theta-\mu)\right\} \qquad \text{for } \theta \in \mathbb{R}^d \\
\text{E}[\,\theta \mid \mu, \Lambda\,] &= \mu \qquad\qquad \text{Var}[\,\theta \mid \mu, \Lambda\,] = \text{E}[\,\theta\theta^T \mid \mu, \Lambda\,] - \mu\mu^T = \Lambda
\end{aligned}$$

$$\begin{aligned}
p(x, y) &= \text{Normal}\left(\begin{bmatrix} x \\ y \end{bmatrix} \;\middle|\; \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{bmatrix}\right) \qquad\qquad \Lambda_{yx} = \Lambda_{xy}^T \\
p(y \mid x) &= \text{Normal}\left(y \mid \mu_y + \Lambda_{yx}\Lambda_{xx}^{-1}(x - \mu_x), \Lambda_{yy} - \Lambda_{yx}\Lambda_{xx}^{-1}\Lambda_{xy}\right)
\end{aligned}$$

$$\begin{aligned}
p(x, y) &= p(y)p(x \mid y) = \text{Normal}\,(y \mid \mu_y, \Lambda_{yy})\,\text{Normal}\,(x \mid Wy + \mu_r, R) \\
p(y \mid x) &= \text{Normal}\left(y \mid \Lambda_{y|x}(\Lambda_{yy}^{-1}\mu_y + W^T R^{-1}(x - \mu_r)), \Lambda_{y|x}\right) \\
\Lambda_{y|x} &= (\Lambda_{yy}^{-1} + W^T R^{-1} W)^{-1}
\end{aligned}$$