Notes on Probabilistic PCA with Missing Values

Jakob Verbeek
INRIA Rhône-Alpes
Updated March 18, 2009

1 Model

Let $\mathbf{y} \in \mathbb{R}^D$ denote a data vector $\mathbf{x} \in \mathbb{R}^d$ denote the vector of principal component coordinates, we let

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, \mathbf{I}), \tag{1}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}; C^\top \mathbf{x}, \sigma^2 \mathbf{I}\right), \tag{2}$$

where C is a $d \times D$ matrix with the projection vectors from the principal component coordinates to the data coordinates. The conditional on $\mathbf{x}$ given $\mathbf{y}$ is then given by

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma), \tag{3}$$
$$\boldsymbol{\mu} = \sigma^{-2} \Sigma C \mathbf{y}, \tag{4}$$
$$\Sigma^{-1} = \mathbf{I} + \sigma^{-2} C C^\top. \tag{5}$$

When only a subset of the coordinates of $\mathbf{y}$ is observed, we replace C above with the C_o which has only the columns corresponding to the observed values, and similar for $\mathbf{y}$ which is replaced by the observed part $\mathbf{y}_o$.

Our goal is now to find the parameters C and σ that maximize the likelihood of some observed data: vectors $\mathbf{y}$ that are fully or partially observed. To do so, we use an EM algorithm that estimates in the E-step the missing values: the vectors $\mathbf{x}$ and the missing parts of the $\mathbf{y}$ which we denote by $\mathbf{y}_h$. In the M-step we fix these estimates, and maximize the expected joint log-likelihood of $\mathbf{x}$ and $\mathbf{y}$.

For simplicity we assume that the distribution over $\mathbf{x}$ and $\mathbf{y}_h$ factors so that we write a lower-bound on the data log-likelihood as

$$\log p(\mathbf{y}_o) \geq \log p(\mathbf{y}_o) - D(q(\mathbf{x})q(\mathbf{y}_h) \| p(\mathbf{x}, \mathbf{y}_h | \mathbf{y}_o)) \tag{6}$$
$$= H(q(\mathbf{x}) + H(q(\mathbf{y}_h)) + \mathbb{E}_q[\log p(\mathbf{x}) + \log p(\mathbf{y}|\mathbf{x})] \tag{7}$$

We will now maximize this bound, in the E-step with respect to the distributions q, and in the M-step with respect to the parameters.

2 E-step

From the above we find the optimal distributions q as

$$q(\mathbf{y}_h) \propto \exp \int q(\mathbf{x}) \log p(z_h|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}_h; C_h^\top \bar{\mathbf{x}}, \sigma^2 \mathbf{I}\right), \tag{8}$$

$$q(\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y}_o) \exp \int q(\mathbf{y}_h) \log p(\mathbf{y}_h|\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; \sigma^{-2} \Sigma C \bar{\mathbf{y}}, \Sigma\right), \tag{9}$$

where $\bar{\mathbf{y}}$ is the mean of $q(\mathbf{y}_h)$ for the missing values and $\mathbf{y}_o$ for the observed part, and $\bar{\mathbf{x}}$ is the mean of $q(\mathbf{x})$.

3 M-step

The expectation in Eq. (7), summed over N data, can be expanded as

$$\sum_{n=1}^{N} \mathbb{E}_{q_n}[\log p(\mathbf{x}_n) + \log p(\mathbf{y}_n|\mathbf{x}_n)] = \tag{10}$$

$$-\frac{ND}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_n \|\bar{\mathbf{y}}_n - C^\top\bar{\mathbf{x}}_n\|^2 - \mathrm{Tr}\{C^\top\Sigma C\}\right) - \frac{D_h}{2\sigma^2}\sigma^2_{\text{old}} \tag{11}$$

$$-\frac{1}{2}\sum_n \|\bar{\mathbf{x}}_n\|^2 - \frac{N}{2}\mathrm{Tr}\{\Sigma\} \tag{12}$$

where D_h denotes the total number of missing values, and σ_{old} is the current value of σ that was used in the E-step to compute the q.

Maximizing this over C and σ we get

$$C = \left(N\Sigma + \bar{X}\bar{X}^\top\right)^{-1}\bar{X}\bar{Y}^\top, \tag{13}$$

$$\sigma^2 = \frac{1}{ND}\left(N\mathrm{Tr}\{C^\top\Sigma C\} + \sum_n \|\bar{\mathbf{y}}_n - C^\top\bar{\mathbf{x}}_n\| + D_h\sigma^2_{\text{old}}\right), \tag{14}$$

where $\bar{X}$ and $\bar{Y}$ denote matrices that collect all $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ as columns.

4 Objective function

Given the expansion of Eq. (7) above, we only need the expression for the entropy of a Gaussian to calculate the EM bound of Eq. (7). It is well known that the entropy for a Gaussian with covariance matrix Σ, for fixed dimensionality, is given up to an additive constant by

$$H = \frac{1}{2}\log|\Sigma|. \tag{15}$$

Combining this with the equations for the parameters found in the M-step, we have that after performing the M-step the bound is calculated as:

$$\log p(\mathbf{y}_o) \geq -\frac{ND}{2}\left(1 + \log\sigma^2\right) - \frac{N}{2}\left(\mathrm{Tr}\{\Sigma\} - \log|\Sigma|\right) - \frac{1}{2}\sum_n \|\mathbf{x}_n\|^2 + \frac{D_h}{2}\log\sigma^2_{\text{old}}.$$

See [1] for the Gaussian identities used in this derivation.

References

[1] S. Roweis. Gausian identities. Online notes, see http://www.cs.toronto.edu/~roweis/notes.html.