Gradient and EM algorithms for PPCA

Xinghu Yao

September 24, 2018

1 Gradient method for PPCA

The log likelihood function of PPCA can be written as

$$
\begin{aligned}
\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2\right) &= \sum_{n=1}^{N} \ln p\left(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2\right) \\
&= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^T \mathbf{C}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}\right).
\end{aligned} \tag{1}
$$

where the $D \times D$ covariance matrix C is defined by

$$
\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{L}. \tag{2}
$$

Setting the derivation w.r.t. $\boldsymbol{\mu}$ equal to zero gives:

$$
\boldsymbol{\mu} = -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \bar{\mathbf{x}}. \tag{3}
$$

The log-likelihood is then simplified as:

$$
\ln p\left(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^T \mathbf{C}^{-1}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right) \tag{4}
$$

or can be written as:

$$
\mathbf{L} \triangleq \ln p\left(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = -\frac{N}{2}\left\{D\ln(2\pi) + \ln|\mathbf{C}| + \operatorname{Tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)\right\} \tag{5}
$$

where S is the data covariance matrix defined by

$$
\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^T \tag{6}
$$

The gradient of the log-likelihood with respect to $\mathbf{W}$ may be obtained from standard matrix differentiation results:

$$
\frac{\partial \mathbf{L}}{\partial \mathbf{W}} = N\left(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}\right). \tag{7}
$$

At the stationary points:

$$
\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W} \tag{8}
$$

By using SVD method and some interesting tricks, we can solve this problem and all solutions of $\mathbf{W}$ can be written as

$$
\mathbf{W}_{ML} = \mathbf{U}_M\left(\mathbf{L}_M - \sigma^2\mathbf{I}\right)^{1/2}\mathbf{R} \tag{9}
$$

where $\mathbf{U}_M$ is a $D \times M$ matrix whose columns are given by any subset of the eigenvectors of the data covariance matrix $\mathbf{S}$, the $M \times M$ diagonal matrix $\mathbf{L}_M$ has elements given by the corresponding eigenvalues λ_i, and $\mathbf{R}$ is an arbitrary $M \times M$ orthogonal matrix. In fact, when the M largest eigenvalues are chosen, the maximum of the likelihood function is obtained. In this case, the columns of $\mathbf{W}$ define the principle subspace of standard PCA and the corresponding maximum likelihood solution for σ^2 is then given by

$$\sigma_{ML}^2 = \frac{1}{D-M} \sum_{i=M+1}^{D} \lambda_i \tag{10}$$

which is the average of the discarded eigenvalues.

2 EM algorithms for PPCA

We first take the expectation of the complete-data log-likelihood w.r.t. the posterior distribution of the latent distribution evaluated using 'old' parameter values. Maximization of this expected complete data log-likelihood then yields the 'new' parameter values. The complete-data log-likelihood function takes the form

$$\ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{W}, \sigma^2\right) = \sum_{n=1}^{N} \left\{ \ln p\left(\mathbf{x}_n | \mathbf{z}_n\right) + \ln p\left(\mathbf{z}_n\right) \right\} \tag{11}$$

where the n^{th} row of the matrix $\mathbf{Z}$ is given by $\mathbf{z}_n$. Recall that $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}), p(\mathbf{x}|\mathbf{z} = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$. Thus, the expectation w.r.t. the posterior distribution over the latent variables can be written as

$$\mathbb{E}\left[\ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)\right] = -\sum_{n=1}^{N} \left\{ \begin{array}{c} \frac{D}{2} \ln\left(2\pi\sigma^2\right) + \frac{1}{2} Tr\left(\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^T\right]\right) \\ + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} \mathbb{E}\left[\mathbf{z}_n\right]^T \mathbf{W}^T \left(\mathbf{x}_n - \boldsymbol{\mu}\right) \\ + \frac{1}{2\sigma^2} Tr\left(\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^T\right] \mathbf{W}^T \mathbf{W}\right) + \frac{M}{2\ln(2\pi)} \end{array} \right\} \tag{12}$$

E-Step: We use the old parameter to evaculate

$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x}_n - \bar{\mathbf{x}}) \tag{13}$$

$$\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] = \sigma^2 \mathbf{M}^{-1} + [\mathbf{z}_n][\mathbf{z}_n]^T \tag{14}$$

This follows directly from

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{x} - \boldsymbol{\mu}\right), \sigma^2\mathbf{M}^{-1}\right), \mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I} \tag{15}$$

together with the standard result

$$\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^T] = \text{cov}[\mathbf{z}] + \mathbb{E}[\mathbf{z}_n] \mathbb{E}[\mathbf{z}_n]^T \tag{16}$$

Substituting Eq. (15) and Eq. (16) into Eq. (12), we can compute the expectation result.
M-Step: We can get the two M-equations by setting the derivatives w.r.t $\mathbf{W}$ and σ^2 to zero, which is

$$\frac{\partial \mathbb{E}\left[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)\right]}{\partial \mathbf{W}} = \sum_{n=1}^{N} \left\{ \frac{1}{\sigma^2} \left(\mathbf{x}_n - \boldsymbol{\mu}\right) \mathbb{E}\left[\mathbf{z}_n\right]^T - \frac{1}{\sigma^2} \mathbf{W} \mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^T\right] \right\} = 0 \tag{17}$$

$$\begin{aligned} \frac{\partial \mathbb{E}\left[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)\right]}{\partial \sigma^2} = \sum_{n=1}^{N} \Bigg\{ & -\frac{D}{2\sigma^2} - \frac{1}{\sigma^4} \mathbb{E}\left[\mathbf{z}_n\right]^T \mathbf{W}^T \left(\mathbf{x}_n - \boldsymbol{\mu}\right) + \\ & \frac{1}{2\sigma^4} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 + \frac{1}{2\sigma^4} \text{Tr}\left(\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^T\right] \mathbf{W}^T \mathbf{W}\right) \Bigg\} = 0 \end{aligned} \tag{18}$$

It is worth to say that we used $\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{A}\mathbf{B}\mathbf{A}^T) = \mathbf{A}(\mathbf{B} + \mathbf{B^T})$, $\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{A}\mathbf{B}) = \mathbf{B}^T$. So, it is clear that we can get the following equations

$$\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{N} \left(\mathbf{x}_n - \bar{\mathbf{x}}\right) \mathbb{E}\left[\mathbf{z}_n\right]^T\right] \left[\sum_{n=1}^{N} \mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^T\right]\right]^{-1} \tag{19}$$

$$\sigma_{\text{new}}^2 = \frac{1}{ND} \sum_{n=1}^{N} \left\{ \|\mathbf{x}_n - \bar{\mathbf{x}}\|^2 - 2\mathbb{E}\left[\mathbf{z}_n\right]^T \mathbf{W}_{new}^T \left(\mathbf{x}_n - \bar{\mathbf{x}}\right) + \text{Tr}\left(\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n^T\right] \mathbf{W}_{new}^T \mathbf{W}_{new}\right) \right\}. \tag{20}$$