Conditional Gaussian and Marginal Gaussian

Xinghu Yao

September 22, 2018

1 Question

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{1}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \tag{2}$$

Try yo derive the following conditional distribution and marginal distribution:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right) \tag{3}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\mathbf{x}_b - \boldsymbol{\mu}_b. \tag{4}$$

$$p(\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \tag{5}$$

2 Conditional Gaussian

According to the definition of condition distribution, we have $p(\mathbf{x}_a|\mathbf{x}_b) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_b)}$. Thus, through fixing $\mathbf{x}_b$ to the observed value and normalizing the resulting expression with $p(\mathbf{x}_b)$, we can get the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$.

We firstly consider the quadratic form in the exponent of the Gaussian distribution give by Eq. (1) and Eq. (2). In fact, we have

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
&= -\frac{1}{2}\begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix} \\
&= -\frac{1}{2}\left[(\mathbf{x}_a - \boldsymbol{\mu}_a)^T, (\mathbf{x}_b - \boldsymbol{\mu}_b)^T\right] \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \begin{pmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{pmatrix} \\
&= -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^T \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&- \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_b)^T \boldsymbol{\Lambda}_{bb}(\mathbf{x}_b - \boldsymbol{\mu}_b).
\end{aligned} \tag{6}$$

We see that as a function of $\mathbf{x}_a$, this is a quadratic form, thus the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is a Gaussian distribution. Noticing that the exponent of a general Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + const \tag{7}$$

where 'const' denotes terms which are independent of $\mathbf{x}$. Consider the functional dependence of Eq. (6) on $\mathbf{x}_b$ in which $\mathbf{x}_a$ is regarded as a constant. The second order of $\mathbf{x}_a$ can be written

$$-\frac{1}{2}\mathbf{x}_a^T \Lambda_{aa}\mathbf{x}_a \tag{8}$$

Comparing Eq. (7) and Eq. (8), we can immediately get the covariance matrix of $p(\mathbf{x}_a|\mathbf{x}_b)$ is given by

$$\boldsymbol{\Sigma}_{a|b} = \Lambda_{aa}^{-1}. \tag{9}$$

Now consider the linear form of $\mathbf{x}_a$ in Eq. (6)

$$\mathbf{x}_a^T \left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right\} \tag{10}$$

Comparing Eq. (7) Eq. (10), we can get $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \mathbf{x}_a^T\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right\}$. Thus, we have

$$\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\right\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)\end{aligned} \tag{11}$$

Combing Eq. (9) and Eq. (11), we can get $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right)$ where $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\mathbf{x}_b - \boldsymbol{\mu}_b$.

3 Marginal Gaussian

In fact, the Marginal Gaussian distribution can be written as

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, d\mathbf{x}_b \tag{12}$$

Picking out those terms only involve $\mathbf{x}_b$ in the joint distribution $p(\mathbf{x}_a, \mathbf{x}_b)$, we have

$$-\frac{1}{2}\mathbf{x}_b^T\boldsymbol{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^T\mathbf{m} = -\frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right)^T\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right) + \frac{1}{2}\mathbf{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m} \tag{13}$$

where $\mathbf{m} = \boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)$ Now we turn to consider the general Gaussian distribution, which is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{14}$$

From Eq. (14) we can see the coefficient of Gaussian distribution is independent of the mean and only governed by the determinant of the covariance matrix. Back to Eq. 13, we can see the integration over $\mathbf{x}$ is as follows

$$\int \exp\left\{-\frac{1}{2}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right)^T\boldsymbol{\Lambda}_{bb}\left(\mathbf{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\mathbf{m}\right)\right\}\mathrm{d}\mathbf{x}_b. \tag{15}$$

This integration is irrelevant with the mean so we can margin out $mathbfx_b$ after the integration.Combing the $\mathbf{x}_b^T\mathbf{m}$ in Eq. (13) with the remaining terms from Eq. (6), we have

$$\begin{aligned}&\frac{1}{2}\left[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)\right]^T\boldsymbol{\Lambda}_{bb}^{-1}\left[\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{ba}\left(\mathbf{x}_a - \boldsymbol{\mu}_a\right)\right] - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^T\left(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b\right) + \text{const} \\ &= \frac{1}{2}\mathbf{x}_a^T\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)\mathbf{x}_a + \mathbf{x}_a^T\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\right)^{-1}\boldsymbol{\mu}_a + \text{const}\end{aligned} \tag{16}$$

where 'const' denotes quantities independent of $\mathbf{x}_a$. Comparing Eq. (16) with Eq. (7), we can get the covariance of $p(\mathbf{x}_a)$ is

$$\boldsymbol{\Sigma}_a = \left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1} = \boldsymbol{\Sigma}_{aa} \tag{17}$$

and the mean of $p(\mathbf{x}_a)$ is given by

$$\boldsymbol{\mu}_a = \boldsymbol{\Sigma}_a\left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)\boldsymbol{\mu}_a \tag{18}$$

We can summarized Eq. (17) and Eq. (18) as

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{19}$$

Bayes' theorem for Gaussian variables

Xinghu Yao

September 22, 2018

1 Question

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \tag{1}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right) \tag{2}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\right) \tag{3}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}\left\{\mathbf{A}^T\mathbf{L}\mathbf{y} - \mathbf{b} + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}\right) \tag{4}$$

where

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)^{-1} \tag{5}$$

2 Solution

First, it is easy (just follow the definition of matrix multiplication) to prove the following equation for the inverse of a partitioned matrix

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{N}^{-1} \end{pmatrix} \tag{6}$$

where we have defined $\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$.
We define $\mathbf{z} = (\mathbf{x}, \mathbf{y})^T$ as the joint distribution over $\mathbf{x}$ and $\mathbf{y}$ and consider the log of $\mathbf{z}$

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}^T)\boldsymbol{\Lambda}(\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^{-1}\mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) + \text{const} \end{aligned} \tag{7}$$

The second order terms in Eq. 7 can be written as

$$\begin{aligned} &-\frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}(\boldsymbol{\Lambda} + \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{A})\boldsymbol{x} - \frac{1}{2}\mathbf{y}^{\mathbf{T}}\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^{\mathbf{T}}\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}\mathbf{L}\mathbf{y} \\ &= -\frac{1}{2}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\mathbf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathbf{T}}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \\ &= -\frac{1}{2}\mathbf{z}^T\mathbf{R}\mathbf{z} \end{aligned} \tag{8}$$

The linear terms in Eq. 7 can be written as

$$\mathbf{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{b} + \mathbf{y}^T\mathbf{L}\mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \mathbf{A}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix}. \tag{9}$$

Thus, the inverse matrix of covariance matrix can be written as

$$\mathbf{R} = \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^{\mathbf{T}}\mathbf{L}\mathbf{A} & -\mathbf{A}^{\mathbf{T}}\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{pmatrix} \tag{10}$$

And using the Eq. 6, we have

$$\text{cov}[\mathbf{z}] = \mathbf{R}^{-1} = \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{\Lambda}^{-1}\mathbf{A}^T \\ \mathbf{A}\mathbf{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T \end{pmatrix}. \tag{11}$$

Similar to the former report, we have

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \mathbf{A}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}. \tag{12}$$

Making use of Eq. 11 and Eq. 12, we can get the mean and covariance of the marginal distribution $p(\mathbf{y})$, which is

$$\mathbb{E}[\mathbf{y}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \tag{13}$$

$$\text{cov}[\mathbf{y}] = \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T. \tag{14}$$

This means $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T)$. By using the last reports results, we have

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = \left(\mathbf{\Lambda} + \mathbf{A}^\mathbf{T}\mathbf{L}\mathbf{A}^{-1}\right)\left\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\right\} \tag{15}$$

$$\text{cov}[\mathbf{x}|\mathbf{y}] = \mathbf{\Lambda} + \mathbf{A}^\mathbf{T}\mathbf{L}\mathbf{A}^{-1}. \tag{16}$$

This means $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\left(\mathbf{\Lambda} + \mathbf{A}^\mathbf{T}\mathbf{L}\mathbf{A}\right)^{-1}\left\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\right\}, \left(\mathbf{\Lambda} + \mathbf{A}^\mathbf{T}\mathbf{L}\mathbf{A}\right)^{-1}\right)$.

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

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

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:

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

or can be written as:

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

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

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

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

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

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

Bayesian Model Comparison

Xinghu Yao

September 25, 2018

1 Question

The Bayesian view of model comparison simply involves the use of probabilities to represent uncertainty in the choice of model, along with a consistent application of the sum and product rules of probability. Consider a data set $\mathcal{D}$ and a set of models $\{\mathcal{M}_i\}$ having parameters $\{\boldsymbol{\theta}_i\}$. For each model we define a likelihood function $p(\mathcal{D}|\boldsymbol{\theta}_i, \mathcal{M}_i)$. If we introduce a prior $p(\boldsymbol{\theta}_i|\mathcal{M}_i)$ for the various models. Try to approximate the distribution as follows using the Laplace Approximation.

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{1}$$

Note that in Eq. (1) each $\mathcal{M}_i$ is omitted to keep the notation uncluttered.

2 The Laplace Approximation

Laplace approximation is a simple but widely used framework which aims to find a Gaussian approximations to a probability density defined over a set of continuous variables. For a given distribution $p(\mathbf{z}) = f(\mathbf{z})/\mathbf{Z}$ defined over an M-dimensional space $\mathbf{z}$. At a stationary point $\mathbf{z}_0$ the gradient $\nabla f(\mathbf{z})$ will vanish. We therefor consider a Taylor expansion of $\ln f(\mathbf{z})$ centred on the mode $\mathbf{z}_0$ so that

$$\ln f(\mathbf{z}) \backsimeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0) \tag{2}$$

Where the $M \times M$ Hessian matrix A is defined by

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0} \tag{3}$$

and ∇ is the gradient operator. Taking the exponential of both sides we obtain

$$f(\mathbf{z}) \backsimeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\} \tag{4}$$

The distribution $q(\mathbf{z})$ is proportional to $f(\mathbf{z})$ and the appropriate normalization coefficient can be found by using the standard result for a normalized multivariate Gaussian, giving

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1}) \tag{5}$$

where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. This Gaussian distribution will be well defined provided its precision matrix, given by A, is positive definite, which implies that the stationary point $\mathbf{z}_0$ must be a local maximum, not a minimum of a saddle point. In fact, we can also obtain an approximation to the normalization Z. Because we have

$$q(\mathbf{z}) \backsimeq p(\mathbf{z}) = \frac{1}{Z}f(\mathbf{z}) \tag{6}$$

Thus, the approximation to the normalization constant Z have the following form

$$Z \backsimeq \frac{f(\mathbf{z})}{q(\mathbf{z})} = f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \tag{7}$$

3 Solution

Identifying $f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $Z = p(\mathcal{D})$. According to Eq. (7), we have

$$\begin{aligned} \ln p(\mathcal{D}) = \ln(Z) &\backsimeq \ln f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \\ &= \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| \end{aligned} \tag{8}$$

where $\boldsymbol{\theta}_{\mathrm{MAP}}$ is the value of $\boldsymbol{\theta}$ at the mode of the posterior distribution, and $\mathbf{A}$ is the Hessian matrix of second derivatives of the negative log posterior

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}|\mathcal{D}) \tag{9}$$

From Eq. (9), we have

$$\begin{aligned} \mathbf{A} &= -\nabla\nabla \ln p(\mathcal{D}|\theta_{\mathrm{MAP}})p(\theta_{MAP}) \\ &= \mathbf{H} - \nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) \end{aligned} \tag{10}$$

and if $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0)$, this becomes

$$\mathbf{A} = \mathbf{H} + \mathbf{V}_0^{-1}. \tag{11}$$

If we assume that the proir is broad or equivalently that the number of data points is large, we can neglect the term $\mathbf{V}_0^{-1}$ compared to $\mathbf{H}$. Using this result, Eq. (8) can be rewritten in the form

$$\ln p(\mathcal{D}) \backsimeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})^T\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \tag{12}$$

We now again invoke the broad prior assumption, allowing us to neglect the second term on the right hand side of Eq. (12). Since we assume i.i.d data, $\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})$ consists of a sum of terms, one term for each datum, and we can consider the following approximation:

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{H}_n = N\widehat{\mathbf{H}} \tag{13}$$

where $\mathbf{H}_n$ is the contribution from the n^{th} data point and

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}_n \tag{14}$$

Combining this with the properties of the determinant, we have

$$\ln|\mathbf{H}| = \ln|N\widehat{\mathbf{H}}| = \ln\left(N^M|\widehat{\mathbf{H}}|\right) = M\ln N + \ln|\widehat{\mathbf{H}}| \tag{15}$$

Where M is the dimensionality of $\boldsymbol{\theta}$. Note that we are assuming that $\widehat{\mathbf{H}}$ has full rank M. Finally, using this result together Eq. (12), we obtain

$$\ln p(\mathcal{D}) \backsimeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}M\ln N \tag{16}$$

Auto-Encoding Variational Bayes

Xinghu Yao

September 26, 2018

1 Problem Establishment

Using auto-encoding variational Bayes we can perform efficient approximate inference and learning with directed probabilistic models whose continuous latent variables and/or parameters have intractable posterior distributions.

Let us consider some dataset $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ consisting of N i.i.d samples of some continuous or discrete variable $\mathbf{x}$. We assume that the data are generated by some random process, involving an unobserved continuous random variable $\mathbf{z}$. The process consists of two steps: (1) a value $\mathbf{z}^{(i)}$ is generated from some prior distribution $p_{\boldsymbol{\theta}}(\mathbf{z})$. (2) a value $\mathbf{x}^{(i)}$ is generated from some conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

We can use two networks to train a generative model and the two parts of our networks have the following relationship.

$$\mathbf{x} \xrightarrow{F_{\boldsymbol{\theta}}} (\mathbf{z}|\mathbf{x}) \xrightarrow{G_{\boldsymbol{\theta}}} \hat{\mathbf{x}} \tag{1}$$

where $F_{\boldsymbol{\theta}}$ is a network to cover the latent hidden variable $\mathbf{z}$ and $G_{\boldsymbol{\theta}}$ is another network to decode $\mathbf{x}$ using the hidden variable $\mathbf{z}$. Thus, the marginal likelihood of this structure can be written as:

$$\begin{aligned}
\ln p(\mathbf{x}) &= \ln \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\
&= \ln \int \frac{p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} p(\mathbf{z}) q(\mathbf{z}|\mathbf{x})\mathrm{d}\mathbf{z} \\
&= \ln \mathop{\mathbb{E}}_{\mathbf{z}\backsim(\mathbf{z}|\mathbf{x})} \left(\frac{p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} p(\mathbf{z}) \right) \\
&\geq \mathop{\mathbb{E}}_{\mathbf{z}\backsim q(\mathbf{z}|\mathbf{x})} \ln \left(\frac{p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} p(\mathbf{z}) \right) && \text{(Jensen Inequality)} \\
&= \mathop{\mathbb{E}}_{\mathbf{z}\backsim q(\mathbf{z}|\mathbf{x})} \ln \frac{p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \mathop{\mathbb{E}}_{\mathbf{z}\backsim q(\mathbf{z}|\mathbf{x})} \ln p(\mathbf{z}) \\
&= \int q(\mathbf{z}|\mathbf{x}) \ln p(\mathbf{x}|\mathbf{z})\mathrm{d}\mathbf{z} - \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right]
\end{aligned} \tag{2}$$

where $q(\mathbf{z}|\mathbf{x})$ is the probabilistic encoder, since given a datapoint $\mathbf{x}$ it produce a distribution over the possible values of the code $\mathbf{z}$ from which the datapoint $\mathbf{x}$ could have been generated. In a similar vein we will refer to $p(\mathbf{x}|\mathbf{z})$ as a probabilistic decoder, since given a code $\mathbf{z}$ it produce a distribution over the possible corresponding values of $\mathbf{x}$. Thus, we can get the following optimization problem

$$\max_{q(\mathbf{z}|\mathbf{x})} \mathop{\mathbb{E}}_{\mathbf{z}\backsim q(\mathbf{z}|\mathbf{x})} \left[\ln p(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right] \tag{3}$$

2 The reparameterization trick

In order to solve our problem we invoke an alternative method for generating samples from $q(\mathbf{z}|\mathbf{x})$. The essential parameterization trick is quite simple. Let $z \backsim p(z|x) = \mathcal{N}(\mu, \sigma^2)$. In this case, a valid reparameterization is $z = \mu + \sigma\epsilon$, where ϵ is an auxiliary noise variable $\epsilon \backsim \mathcal{N}(0, 1)$