

The evaluation of this conditional distribution can be seen as an example of Bayes' theorem, in which we interpret  $p(\mathbf{x})$  as a prior distribution over  $\mathbf{x}$ . If the variable  $\mathbf{y}$  is observed, then the conditional distribution  $p(\mathbf{x}|\mathbf{y})$  represents the corresponding posterior distribution over  $\mathbf{x}$ . Having found the marginal and conditional distributions, we have effectively expressed the joint distribution  $p(\mathbf{z}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$  in the form  $p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$ .

These results can be summarized as follows. Given a marginal Gaussian distribution for  $\mathbf{x}$  and a conditional Gaussian distribution for  $\mathbf{y}$  given  $\mathbf{x}$  in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (3.97)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}), \quad (3.98)$$

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

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (3.99)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (3.100)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \quad (3.101)$$

### 3.2.7 Maximum likelihood

Given a data set  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^T$  in which the observations  $\{\mathbf{x}_n\}$  are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by maximum likelihood. The log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}). \quad (3.102)$$

By simple rearrangement, we see that the likelihood function depends on the data set only through the two quantities

$$\sum_{n=1}^N \mathbf{x}_n, \quad \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T. \quad (3.103)$$

These are known as the *sufficient statistics* for the Gaussian distribution. Using (A.19), the derivative of the log likelihood with respect to  $\boldsymbol{\mu}$  is given by

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}), \quad (3.104)$$

and setting this derivative to zero, we obtain the solution for the maximum likelihood estimate of the mean:

$$\boldsymbol{\mu}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \quad (3.105)$$