

**6.23** Here, the difference from the previous results is that the target is also a vector.

We take account of the noise on the observed target values:

$$\mathbf{t}_n = \mathbf{y}_n + \boldsymbol{\epsilon}_n$$

where  $\mathbf{y}_n = \mathbf{y}(\mathbf{x}_n)$ .

Assuming that the noise is independent for each observation and each target vector element,

$$p(\mathbf{t}_n | \mathbf{y}_n) = \mathcal{N}(\mathbf{t}_n | \mathbf{y}_n, \mathbf{B}^{-1})$$

where  $\mathbf{B} \in \mathbb{R}^{D \times D}$  is a diagonal matrix of precisions for the individual components of the target vectors.

To model the joint distribution of the target values  $\mathbf{t}_1, \dots, \mathbf{t}_N$  conditioned on the values of  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ , we can vectorize the outputs such that:

$$\mathbf{t} = \begin{bmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \vdots \\ \mathbf{t}_N \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} \quad \mathbf{D}^{-1} = \begin{bmatrix} \mathbf{B}^{-1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{-1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{B}^{-1} \end{bmatrix}$$

$$p(\mathbf{t} | \mathbf{y}) = \mathcal{N}(\mathbf{t} | \mathbf{y}, \mathbf{D}^{-1})$$

The Gram Matrix here will look like:

$$\mathbf{K}_{\text{full}} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) \mathbf{I}_D & k(\mathbf{x}_1, \mathbf{x}_2) \mathbf{I}_D & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \mathbf{I}_D \\ k(\mathbf{x}_2, \mathbf{x}_1) \mathbf{I}_D & k(\mathbf{x}_2, \mathbf{x}_2) \mathbf{I}_D & \dots & k(\mathbf{x}_2, \mathbf{x}_N) \mathbf{I}_D \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) \mathbf{I}_D & k(\mathbf{x}_N, \mathbf{x}_2) \mathbf{I}_D & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \mathbf{I}_D \end{bmatrix}$$

where each  $D \times D$  submatrix of  $\mathbf{K}_{\text{full}}$  has the form:

$$k(\mathbf{x}_i, \mathbf{x}_j) \mathbf{I}_D = \begin{bmatrix} k(\mathbf{x}_i, \mathbf{x}_j) & 0 & \dots & 0 \\ 0 & k(\mathbf{x}_i, \mathbf{x}_j) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & k(\mathbf{x}_i, \mathbf{x}_j) \end{bmatrix}$$

This structure arises because the output dimensions of the target vector  $\mathbf{t}_n \in \mathbb{R}^D$  are assumed to be conditionally independent given the inputs. As a result, the covariance between different output dimensions (i.e., the off-diagonal entries) is zero, and each output dimension contributes only to its own diagonal entry in the block.

This can be written more simply using Kronecker Product:

$$\mathbf{K}_{\text{full}} = \mathbf{K} \otimes \mathbf{I}_D$$

where:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & k(\mathbf{x}_N, \mathbf{x}_2) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

$$\implies p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K}_{\text{full}})$$

Using 6.61 and 6.62, the marginal distribution of  $\mathbf{t}$  is given by:

$$p(\mathbf{t}) = \mathcal{N}(\mathbf{0}, \mathbf{C})$$

where

$$\begin{aligned} \mathbf{C} &= \mathbf{K}_{\text{full}} + \mathbf{D}^{-1} \\ &= \mathbf{K} \otimes \mathbf{I}_D + \mathbf{I}_N \otimes \mathbf{B}^{-1} \end{aligned}$$

From (6.61), the joint distribution over  $\mathbf{t}_1, \dots, \mathbf{t}_{N+1}$  will be given by :

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1} | \mathbf{0}, \mathbf{C}_{N+1})$$

where:

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{E} \\ \mathbf{E}^T & \mathbf{F} \end{pmatrix}$$

and

$$\mathbf{E} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_{N+1})\mathbf{I}_D \\ k(\mathbf{x}_2, \mathbf{x}_{N+1})\mathbf{I}_D \\ \vdots \\ k(\mathbf{x}_N, \mathbf{x}_{N+1})\mathbf{I}_D \end{bmatrix} \quad \mathbf{F} = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1})\mathbf{I}_D + \mathbf{B}^{-1}$$

Using the results (2.81) and (2.82), we see that the conditional distribution  $p(\mathbf{t}_{N+1}|\mathbf{t})$  is a Gaussian distribution with mean and covariance:

$$\boldsymbol{\mu}(\mathbf{x}_{N+1}) = \mathbf{E}^T \mathbf{C}_N^{-1} \mathbf{t}$$

$$\boldsymbol{\Sigma}(\mathbf{x}_{N+1}) = \mathbf{F} - \mathbf{E}^T \mathbf{C}_N^{-1} \mathbf{E}$$