7.2. PARAMETER ESTIMATION

The parameters of the multivariate normal distribution are the vector of feature means μ and the variance–covariance matrix Σ. The estimate of μ from a set of feature vectors is the vector of sample means $\bar{\mathbf{x}}^T = (\bar{x}_1, \ldots, \bar{x}_d)$, where d is the number of features, $\bar{x}_i = (1/N) \Sigma_{k=1}^{N} x_{ik}$, and N is the number of objects. The estimate of Σ is the sample variance–covariance matrix S. The elements of S are the sample variances $s_{ii} = [1/(N-1)] \Sigma_{k=1}^{N} (x_{ik} - \bar{x}_i)^2$ and the sample covariances $s_{ij} = [1/(N-1)] \Sigma_{k=1}^{N} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$. We use the factor $1/(N-1)$ to obtain an unbiased estimate, as we explained in the last chapter. Also, the estimated correlation between feature i and feature j is given by $r = s_{ij}/\sqrt{s_{ii}s_{jj}}$.

In the general case, $s_{ij} \neq 0$, if we assume multivariate normality, we need to find the d means of the x_i and the d^2 s_{ij}. But since $s_{ij} = s_{ji}$, there are only $d(d+3)/2$ parameters.