5. Discriminant analysis

We continue from Bayes's rule presented in Section 3 on p. 88

$$P(c_i \mid \mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|c_i)P(c_i)}{\sum_{j=1}^{C} p(\mathbf{x}|c_j)P(c_j)}, i = 1,2,\ldots,C \quad (5.1)$$

where c_i is a class, $\mathbf{x}$ is a p-dimensional vector (data case) and we use class conditional probability (density function) and a priori class probability both in numerator and denominator. See Fig. 5.1

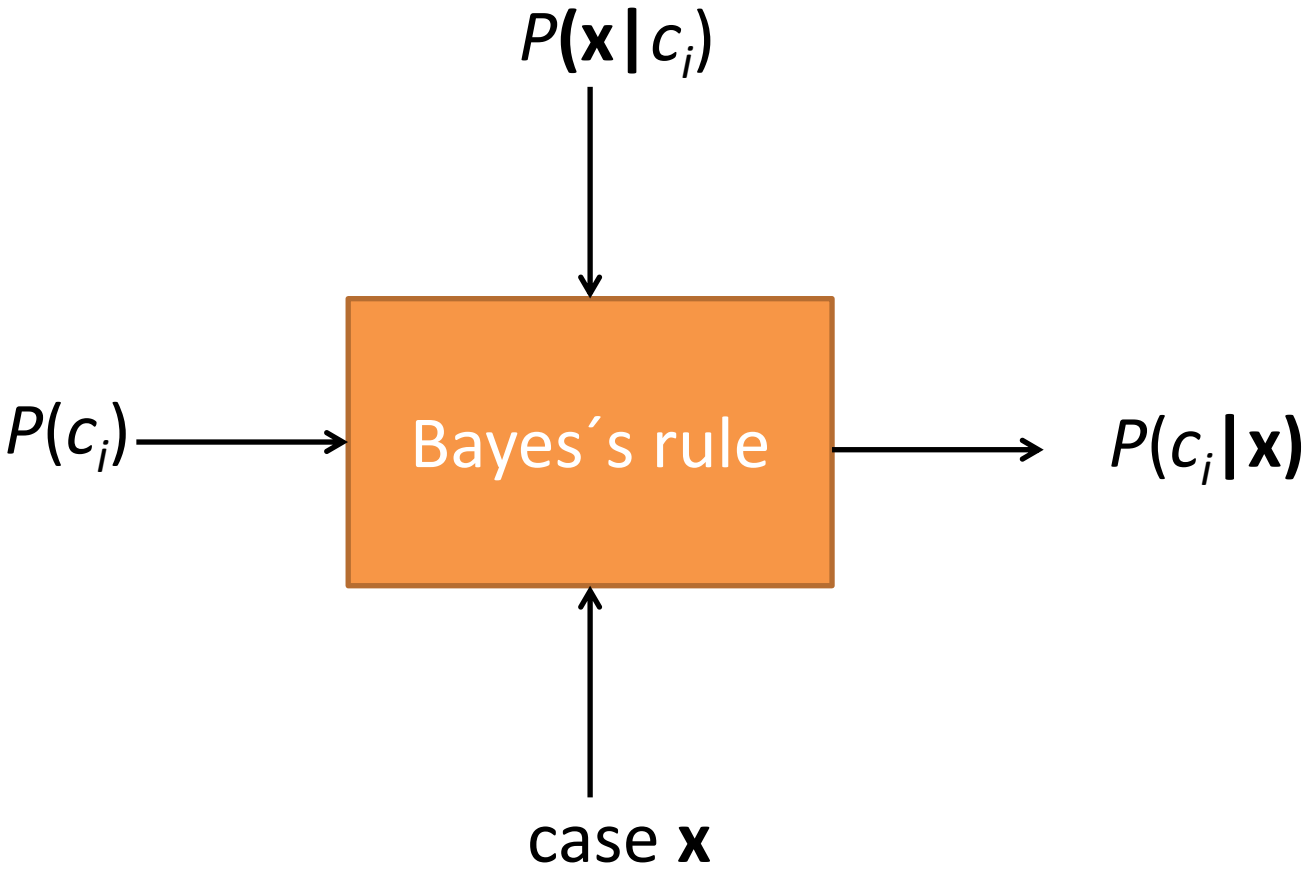

Fig 5.1 Converting a priori class probability to a posteriori probability via vector **x**.

On the basis of Section 3 p. 97-99, we know that minimizing the risk or error probability is equivalent to partitioning the variable space into M regions, classes for classification. If regions R_i and R_j happen to be contiguous, they are separated by *decision boundary* or *decision surface* in the multidimensional variable space. For the minimum error probability, this is described by the following equation.

$$P(c_i \mid \mathbf{x}) - P(c_j \mid \mathbf{x}) = 0$$

From the one side of the boundary this difference is positive and from the other it is negative. Sometimes, instead of working directly with probabilities (or risk functions), it may be more convenient to work with their equivalent function, for example:

$$g_i(\mathbf{x}) = f\left(P\left(c_i \middle| \mathbf{x}\right)\right) \qquad (5.2)$$

Now $f(\cdot)$ is a monotonically increasing function, and $g_i(\mathbf{x})$ is known as a *discriminant function*. The decision test (5.2) is stated

$$\text{classify } \mathbf{x} \text{ in } c_i \text{ if } g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ for all } j \neq i \quad (5.3)$$

and decision boundaries separating contiguous regions are defined in the following way.

$$g_{ij}(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0,\ i,j=1,2,\ldots,C,\ j \neq i$$

Next we focus on a particular family of decision boundaries associated with Bayesian classification and also Gaussian density (normal distribution).

We employ the Gaussian or normal density function from Section 3 p. 117

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) \right) \quad (5.4)$$

where $\boldsymbol{\Sigma}$ is $p \times p$ covariance matrix and $|\boldsymbol{\Sigma}|$ its determinant.

5.1 Quadratic discriminant analysis

Let us define a discriminant function for the *i*th class from (5.1):

$$g_i(\mathbf{x}) = P(c_i|\,\mathbf{x})$$

Given a vector **x** of *p* variables, classification according to (5.1) and (5.3) is based on detecting the greatest discriminant function value. Denominator $p(\mathbf{x})$ is neglected, because it can be reduced as a positive constant for any class c_i for a fixed **x**. (We could even simplify assuming equal a priori probabilities, this would mean choosing the class for which $p(\mathbf{x}|c_i)$ is greatest.)

As written above, any monotonically increasing function $g_i(\mathbf{x})$ is also a valid discriminant function. The logarithmic function meets this requirement and gives an alternative discriminant function:

$$d_i(\mathbf{x}) = \ln\left(P(c_i \mid \mathbf{x})\right)$$

We now simplify the arrangement to illustrate the solution approach.

We attain the following form

$$d_i(\mathbf{x}) = \ln P\left(\mathbf{x}\middle|c_i\right) + \ln P(c_i), i = 1,2,\ldots,C \qquad (5.5)$$

where case **x** follows a multidimensional normal distribution in every class. From (5.4) we derive

$$p(\mathbf{x}\,|\,c_i) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right), \qquad (5.6)$$

$$i = 1,2,\ldots,C$$

where mean vector $\boldsymbol{\mu}_i$ of the *i*th class and covariance matrix and its determinant are used as usual.

By substituting (5.6) into (5.5) we obtain the following form.

$$d_i(\mathbf{x}) = \ln(2\pi)^{-p/2} + \ln\left|\boldsymbol{\Sigma}_i\right|^{-1/2} - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right) + \ln P(c_i), \quad (5.7)$$

$$i = 1,2,\ldots,C$$

The first term is dropped out, because it is the same constant for every class.

In discriminant analysis a test case **x** is classified into class c_i that gives the greatest $d_i(\mathbf{x})$.

Let us consider a simple two-dimensional example and assume that the covariance matrix is as follows.

$$\Sigma_i = \begin{pmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{pmatrix}$$

The (5.7) becomes

$$d_i(\mathbf{x}) = \ln(2\pi)^{-p/2} + \ln\left|\boldsymbol{\Sigma}_i\right|^{-1/2} - \frac{1}{2\sigma_i^2}(x_1^2 + x_2^2)$$

$$+ \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln P(c_i)$$

and obviously the associated decision curves

$$d_i(\mathbf{x})\text{-}d_j(\mathbf{x})=0$$

are quadrics, that is, ellipsoids, parabolas, hyperbolas, or pair of lines.

Fig. 5.2 shows the decision curves corresponding to $P(c_1)=P(c_2)=1/2$, $\boldsymbol{\mu}_1=(0\ 0)^T$ and $\boldsymbol{\mu}_2=(1\ 0)^T$. The covariance matrices are

$$\Sigma_1=\begin{pmatrix}0.1 & 0.0\\ 0.0 & 0.15\end{pmatrix}\quad \Sigma_2=\begin{pmatrix}0.2 & 0.0\\ 0.0 & 0.25\end{pmatrix}$$

for Fig. 5.2(a) and

$$\Sigma_1=\begin{pmatrix}0.1 & 0.0\\ 0.0 & 0.15\end{pmatrix}\quad \Sigma_2=\begin{pmatrix}0.15 & 0.0\\ 0.0 & 0.1\end{pmatrix}$$

for Fig. 5.2(b).

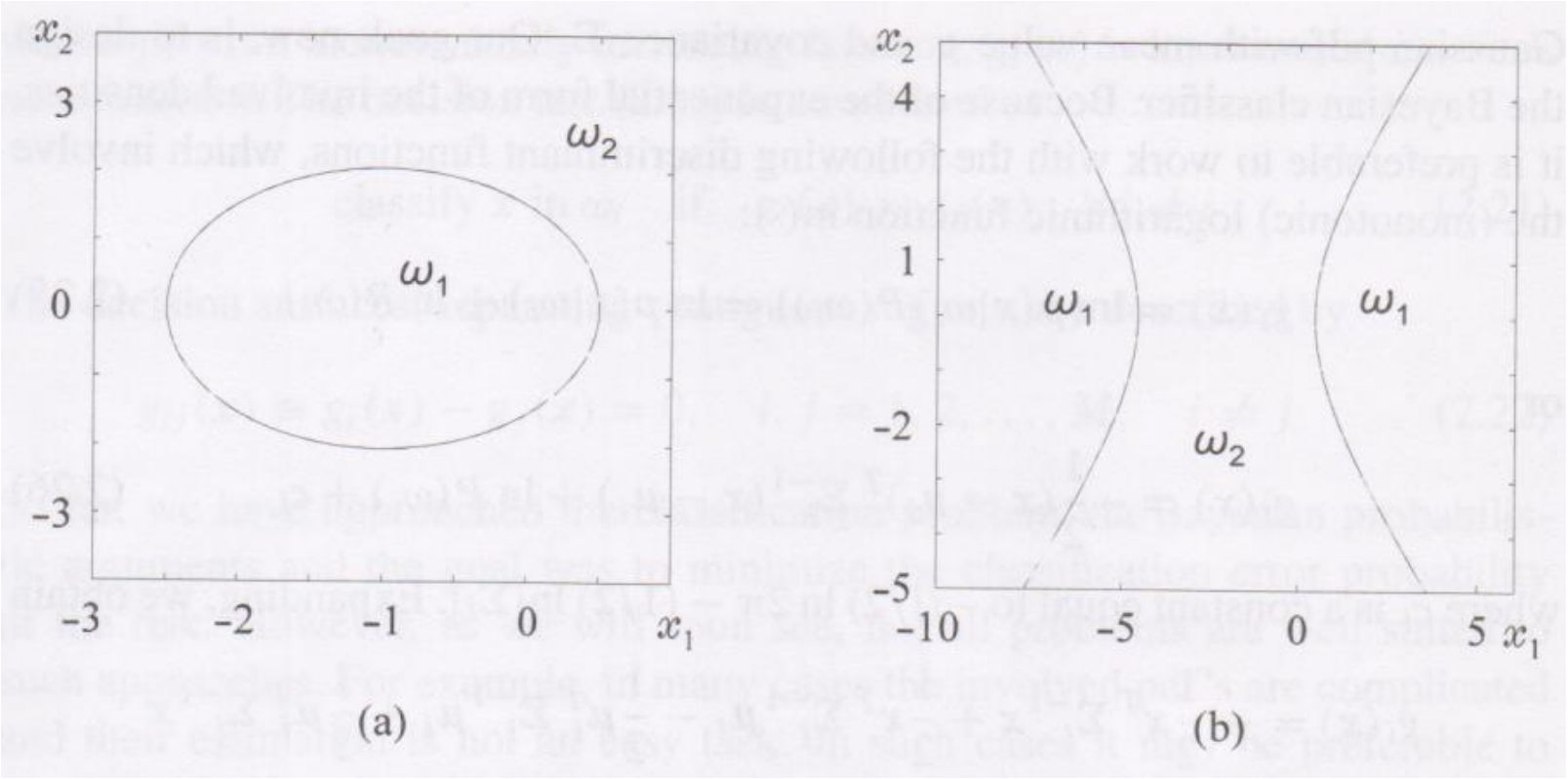

Fig. 5.2 Quadric decision curves, (a) ellipsoid (b) hyperbola (ω_1 and ω_2 are classes).

5.2 Linear discriminant analysis

The only quadratic contribution in (5.7) comes from the term $\mathbf{x}^T \Sigma_i^{-1}\mathbf{x}$. If we now assume that the covariance matrix is the same in all classes, that is, $\Sigma_i = \Sigma$, the quadratic term will be the same in all discriminant functions. Hence, it does not enter into the comparison for computing the maximum and it cancels out in the decision surface equations. Thus, it can be omitted and we redefine

$$d_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \qquad (5.8)$$

where

$$\mathbf{w}_i = \boldsymbol{\mu}_i^{\mathrm{T}} \Sigma^{-1}$$

$$w_{i0} = \ln P(c_i) - \frac{1}{2} \boldsymbol{\mu}_i^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu}_i$$

Hence $d_i(\mathbf{x})$ is a *linear discriminant function* of **x** and the respective decision surfaces are hyperplanes.

For diagonal covariance matrix with equal elements there is $\Sigma=\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the p-dimensional identity matrix and σ^2 variance. Then (5.8) becomes

$$d_i(\mathbf{x}) = \frac{1}{\sigma^2}\boldsymbol{\mu}_i^{\mathrm{T}}\mathbf{x} + w_{i0}$$

Thus, the corresponding decision hyperplanes can be written as

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x}) - d_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_0) = 0$$

where

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \sigma^2 \ln\left(\frac{P(c_i)}{P(c_j)}\right) \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \quad (5.9)$$

where

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + \ldots + a_p^2}$$

is the Euclidean norm of a vector. Now the decision surface is *hyperplane* passing through point $\mathbf{x}_0$. If $P(c_i)=P(c_j)$, then $\mathbf{x}_0=(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)/2$, and the hyperplane passes through the mean of mean vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$.

The geometry is illustrated in Fig. 5.3 for the two-dimensional case. The decision hyperplane (line here) is orthogonal to $\boldsymbol{\mu}_i$-$\boldsymbol{\mu}_j$. For any point $\mathbf{x}$ lying on the decision hyperplane, the vector $\mathbf{x}$-$\mathbf{x}_0$ also lies on the hyperplane.

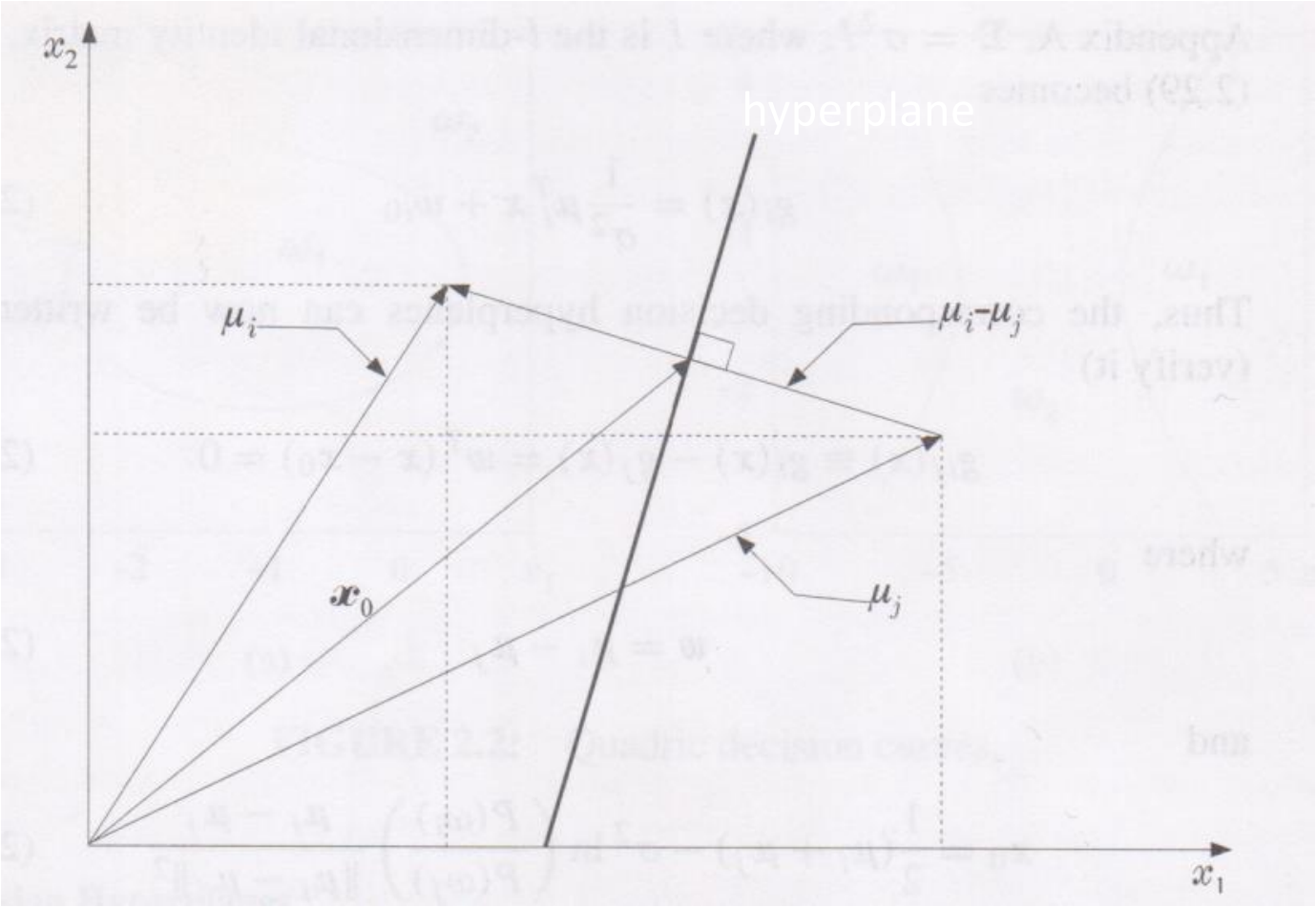

Fig. 5.3 Decision line for two classes and normally distributed vectors with $\Sigma=\sigma^2\mathbf{I}$.

If $P(c_i)<P(c_j)$ (or $P(c_i)>P(c_j)$), the hyperplane is closer to $\boldsymbol{\mu}_i$ (or $\boldsymbol{\mu}_j$). If σ^2 is small with respect to the Euclidean norm of $\boldsymbol{\mu}_i$-$\boldsymbol{\mu}_j$, the location of the hyperplane is rather insensitive to the values of $P(c_i)$ and $P(c_j)$. Small variance indicates that random vectors are clustered within a small radius around the means. Fig. 5.4 illustrates this. For each class, the circles show regions in which cases have a high probability, say 98%, of being found.

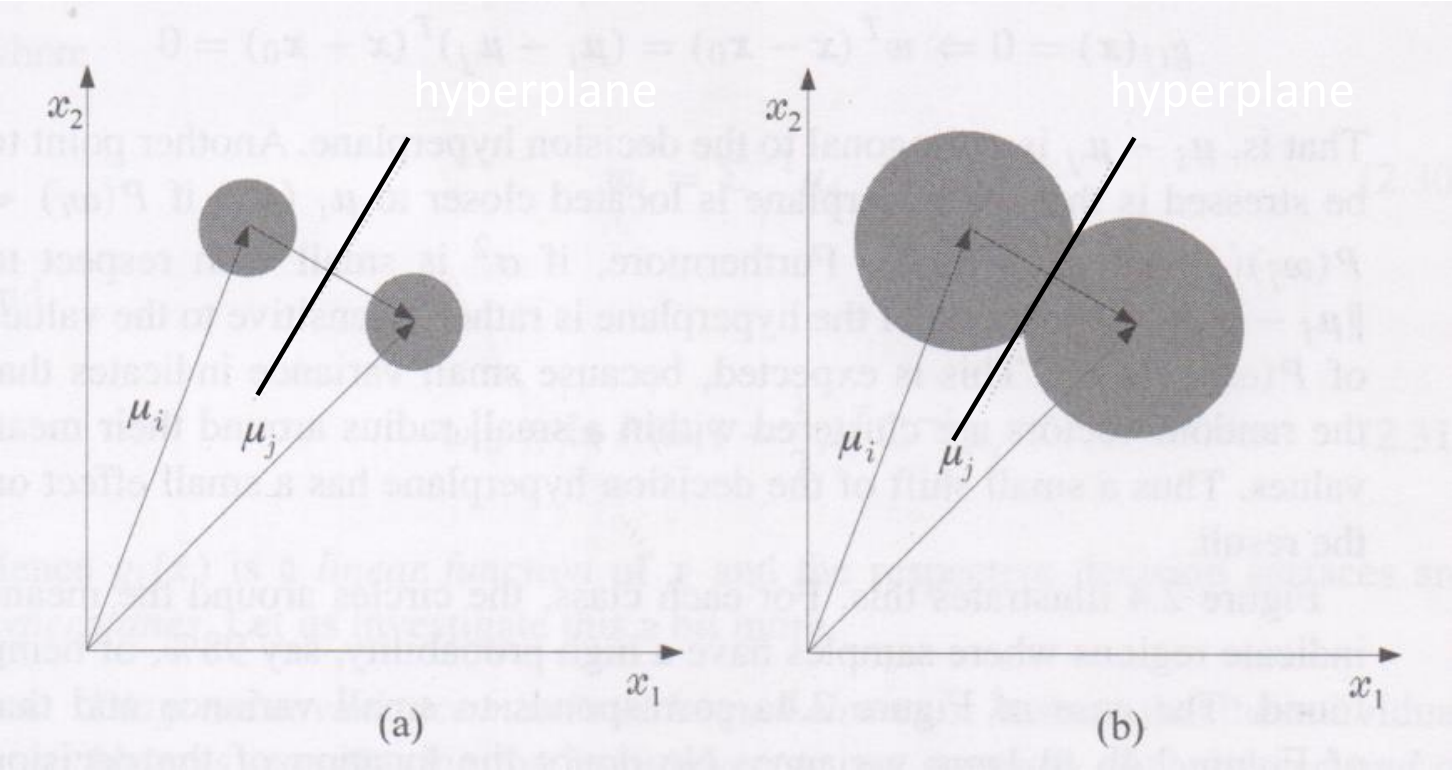

Fig. 5.4 The decision lines for (a) small variance and (b) for large variance. The location of the hyperplane in the latter is much more critical than in the former.

For a nondiagonal covariance matrix we end up with hyperplanes given by

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x}) - d_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}(\mathbf{x} - \mathbf{x}_0) = 0$$

where

$$\mathbf{w} = \Sigma^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)$$

and then the norm from (5.9) is replaced with Mahalanobis distance (norm)

$$\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_{\Sigma^{-1}} = \sqrt{\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)}$$

This is like in the case of the diagonal covariance matrix, with one exception. The decision hyperplane is no longer orthogonal in the vector $\boldsymbol{\mu}_i$-$\boldsymbol{\mu}_j$, but to its linear transformation $\Sigma^{-1}(\boldsymbol{\mu}_i$-$\boldsymbol{\mu}_j)$.

These are illustrated in Fig. 5.5.

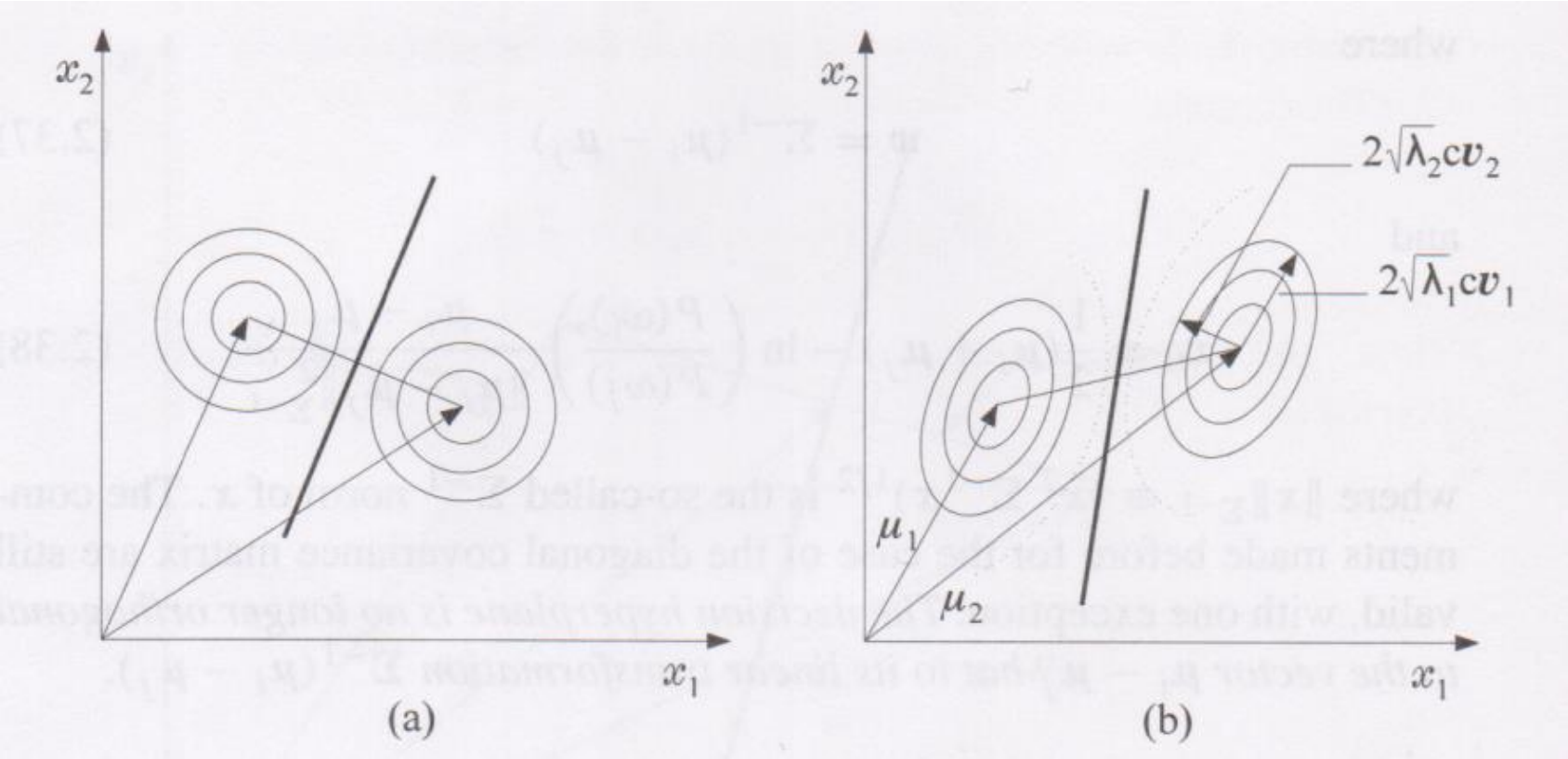

Fig. 5.5 Curves of (a) equal Euclidean distance and (b) equal Mahalanobis distance from the mean points of each class. (λ's and *v*'s are eigenvalues and eigenvectors and *c* is a constant.)

5.3 Logistic discrimination

The basic assumption is that the difference between the logarithms of the class-conditional density functions is linear in the variables of **x**:

$$\ln \frac{p(\mathbf{x} \mid c_1)}{p(\mathbf{x} \mid c_2)} = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}$$

The assumption is satisfied by many families of distributions and thus applicable to a wide range of real data sets depart from normal distribution.

Since the sum of the previous class-conditional probabilities has to be equal to 1, we obtain

$$p(c_2 \mid \mathbf{x}) = \frac{1}{1+\exp\left(\beta_0' + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)}$$

$$p(c_1 \mid \mathbf{x}) = \frac{\exp\left(\beta_0' + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)}{1+\exp\left(\beta_0' + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)} \qquad (5.11)$$

where

$$\beta_0' = \beta_0 + \ln\left(P(c_1)/P(c_2)\right)$$

Discrimination between two classes depends on the ratio $p(c_1|\mathbf{x})/p(c_2|\mathbf{x})$:

Assign to c_1 if $p(c_1|\mathbf{x})/p(c_2|\mathbf{x})>1$, otherwise to c_2.

Using (5.11) we see that the decision about discrimination is determined solely by the linear function as follows:

Assign to c_1 if $\beta_0' + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} > 0$, otherwise to c_2.

Multiclass logistic discrimination

In the multiclass discrimination (regression) problem, the basic assumption is that, for *C* classes

$$\ln \frac{p(\mathbf{x}|c_i)}{p(\mathbf{x}|c_j)} = \beta_{i0} + \boldsymbol{\beta}_i{}^{\mathrm{T}}\mathbf{x}, \quad i, j = 1,2.,\ldots,C, \; i \neq j$$

That log-likelihood ratio is linear for any pair of likelihoods. Again, we may show that the posterior probabilities are of the form

$$p(c_i \mid \mathbf{x}) = \frac{\exp\left(\beta_{i0}' + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}\right)}{1 + \sum_{i=1, i\neq j}^{C} \exp\left(\beta_{i0}' + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}\right)}, i = 1, \ldots, C, i \neq j$$

$$p(c_j \mid \mathbf{x}) = \frac{1}{1 + \sum_{i=1, i\neq j}^{C} \exp\left(\beta_{i0}' + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}\right)}$$

where

$$\beta_{i0}' = \beta_{i0} + \ln\left(P(c_i) / P(c_j)\right)$$

Also, the decision rule about discrimination depends solely on the linear functions:

Assign **x** to c_m if $\max_{\substack{i=1,..,C \\ i\neq j}} \left\{\beta_{i0}' + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}\right\} = \beta_{m0}' + \boldsymbol{\beta}_m^{\mathrm{T}}\mathbf{x} > 0$, otherwise assign **x** to c_j.

Example 1: Demographic vs. crime variables

Table 5.1 Classification accuracies (%) of discriminant analysis after scaling or without it.

Method	Not scaled	Scaled into [0,1]	Standardized
Linear	83.9	83.9	83.9
Logistic	73.2	73.2	73.2

Returning to the example [10] from p. 122 and p. 143, we obtained result that were either worse or better than those given by naive Bayes, and either worse or equal than those given by the nearest neighbor searching.

[10] X. Li, H. Joutsijoki, J. Laurikkala and M. Juhola: Crime vs. demographic factors: application of data mining methods, Webology, Article 132, 12(1), 1-19, 2015.

Example 2: Vertigo data

Let us view Vertigo data[11] from Section 3, p. 124-126, and Section 4, p. 146. For linear discriminant analysis we obtained high accuracy of 90.4%. This was slightly better than the accuracy in Section 4.

Applying Mahalanobis or quadratic discriminant function was not successful because of not positive-definitive matrices obtained in the computation which prevented their use.

(In Matlab classification using linear and quadratic discriminant functions can be made with ’classify’. Logistic discriminant function is ’mnrfit’ after the use of which ’mnrval’ gives probabilities class by class.)

[11] M. Juhola: Data Classification, Encyclopedia of Computer Science and Engineering, ed. B. Wah, John Wiley & Sons, 2008 (print version, 2009, Hoboken, NJ), 759-767.