Recognizing that \(|\mathbf{e}| = 1\), partially differentiating with respect to \(a_i\), and setting the derivative to zero, we obtain

\[ a_i = \mathbf{e}'(x_i - \mathbf{m}). \]  

(83)

Geometrically, this result merely says that we obtain a least-squares solution by projecting the vector \(x_i\) onto the line in the direction of \(\mathbf{e}\) that passes through the sample mean.

This brings us to the more interesting problem of finding the best direction \(\mathbf{e}\) for the line. The solution to this problem involves the so-called scatter matrix \(\mathbf{S}\) defined by

\[ \mathbf{S} = \sum_{i=1}^{n}(x_i - \mathbf{m})(x_i - \mathbf{m})' \]  

(84)

The scatter matrix should look familiar—it is merely \(n - 1\) times the sample covariance matrix. It arises here when we substitute \(a_i\) found in Eq. 83 into Eq. 82 to obtain

\[ J_1(\mathbf{e}) = \sum_{i=1}^{n} a_i^2 - 2 \sum_{i=1}^{n} a_i \mathbf{e}'(x_i - \mathbf{m}) + \sum_{i=1}^{n} ||x_i - \mathbf{m}||^2 \]

\[ = -\mathbf{e}'\mathbf{S}\mathbf{e} - \sum_{i=1}^{n} ||x_i - \mathbf{m}||^2 \]

\[ = -\mathbf{e}'\mathbf{S}\mathbf{e} + \sum_{i=1}^{n} ||x_i - \mathbf{m}||^2. \]  

(85)

Clearly, the vector \(\mathbf{e}\) that minimizes \(J_1\) also maximizes \(\mathbf{e}'\mathbf{S}\mathbf{e}\). We use the method of Lagrange multipliers (described in Section A.3 of the Appendix) to maximize \(\mathbf{e}'\mathbf{S}\mathbf{e}\) subject to the constraint that \(|\mathbf{e}| = 1\). Letting \(\lambda\) be the undetermined multiplier, we differentiate

\[ u = \mathbf{e}'\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}'\mathbf{e} - 1). \]  

(86)

with respect to \(\mathbf{e}\) to obtain

\[ \frac{\partial u}{\partial \mathbf{e}} = 2\mathbf{S}\mathbf{e} - 2\lambda \mathbf{e}. \]  

(87)

Setting this gradient vector equal to zero, we see that \(\mathbf{e}\) must be an eigenvector of the scatter matrix:

\[ \mathbf{S}\mathbf{e} = \lambda \mathbf{e}. \]  

(88)

In particular, because \(\mathbf{e}'\mathbf{S}\mathbf{e} = \lambda\mathbf{e}'\mathbf{e} = \lambda\), it follows that to maximize \(\mathbf{e}'\mathbf{S}\mathbf{e}\), we want to select the eigenvector corresponding to the largest eigenvalue of the scatter matrix. In other words, to find the best one-dimensional projection of the data (best in the least-

sum-of-squared-error sense), we project the data onto a line through the sample mean in the direction of the eigenvector of the scatter matrix having the largest eigenvalue.

This result can be readily extended from a one-dimensional projection to a \(d'\)-dimensional projection. In place of Eq. 81, we write

\[ \mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i, \]  

(89)

where \(d' \leq d\). It is not difficult to show that the criterion function

\[ J_{d'} = \sum_{i=1}^{d'} \left( ||\mathbf{m} + \sum_{j=1}^{d'} a_j \mathbf{e}_j - x_i||^2 \right) \]  

(90)

is minimized when the vectors \(\mathbf{e}_1, \ldots, \mathbf{e}_{d'}\) are the \(d'\) eigenvectors of the scatter matrix having the largest eigenvalues. Because the scatter matrix is real and symmetric, these eigenvectors are orthogonal. They form a natural set of basis vectors for representing any feature vector \(\mathbf{x}\). The coefficients \(a_i\) in Eq. 89 are the components of \(\mathbf{x}\) in that basis, and are called the principal components. Geometrically, if we picture the data points \(x_1, \ldots, x_n\) as forming a \(d'\)-dimensional, hyperellipsoidally shaped cloud, then the eigenvectors of the scatter matrix are the principal axes of that hyperellipsoid. Principal component analysis reduces the dimensionality of feature space by restricting attention to those directions along which the scatter of the cloud is greatest.

3.8.2 Fisher Linear Discriminant

Although PCA finds components that are useful for representing data, there is no reason to assume that these components must be useful for discriminating between data in different classes. If we pool all of the samples, the directions that are discarded by PCA might be exactly the directions that are needed for distinguishing between classes. For example, if we had data for the printed uppercase \(O\) and \(Q\), PCA might discover the gross features that characterize \(O\) and \(Q\), but might ignore the tail that distinguishes an \(O\) from a \(Q\). Where PCA seeks directions that are efficient for representation, discriminant analysis seeks directions that are efficient for discrimination.

We begin by considering the problem of projecting data from \(d\) dimensions onto a line. Of course, even if the samples formed well-separated, compact clusters in \(d\)-space, projection onto an arbitrary line will usually produce a confused mixture of samples from all of the classes and thus produce poor recognition performance. However, by moving the line around, we might be able to find an orientation for which the projected samples are well separated. This is exactly the goal of classical discriminant analysis.

Suppose that we have a set of \(n\) \(d\)-dimensional samples \(x_1, \ldots, x_n\) in the subset \(D_1\) labeled \(\omega_1\) and \(n_2\) in the subset \(D_2\) labeled \(\omega_2\). If we form a linear combination of the components of \(\mathbf{x}\), we obtain the scalar dot product

\[ y = \mathbf{w}' \mathbf{x}, \]  

(91)

and a corresponding set of \(n\) samples \(y_1, \ldots, y_n\) divided into the subsets \(Y_1\) and \(Y_2\). Geometrically, if \(||\mathbf{w}|| = 1\), each \(y_i\) is the projection of the corresponding \(x_i\) onto
a line in the direction of \( w \). Actually, the magnitude of \( w \) is of no real significance, because it merely scales \( y \). The direction of \( w \) is important, however. If we imagine that the samples labeled \( a_1 \) fall more or less into one cluster while those labeled \( a_2 \) fall in another, we want the projections falling onto the line to be well separated, not thoroughly intermingled. Figure 3.5 illustrates the effect of choosing two different values for \( w \) for a two-dimensional example. It should be abundantly clear that if the original distributions are multimodal and highly overlapping, even the "best" \( w \) is unlikely to provide adequate separation, and thus this method will be of little use.

We now turn to the matter of finding the best such direction \( w \), one we hope will enable accurate classification. A measure of the separation between the projected points is the difference of the sample means. If \( m_i \) is the \( d \)-dimensional sample mean given by

\[
 m_i = \frac{1}{n_i} \sum_{x \in S_i} x, \tag{92}
\]

then the sample mean for the projected points is given by

\[
 \bar{\tilde{m}}_i = \frac{1}{n_i} \sum_{y \in \tilde{S}_i} y = \frac{1}{n_i} \sum_{x \in \tilde{S}_i} w^T x = w^T m_i \tag{93}
\]

and is simply the projection of \( m_i \).

It follows that the distance between the projected means is

\[
|\bar{\tilde{m}}_1 - \bar{\tilde{m}}_2| = |w^T (m_1 - m_2)| \tag{94}
\]

and that we can make this difference as large as we wish merely by scaling \( w \). Of course, to obtain good separation of the projected data we really want the difference between the means to be large relative to some measure of the standard deviations for each class. Rather than forming sample variances, we define the scatter for projected samples labeled \( a_1 \) by

\[
 \tilde{z}_i^2 = \sum_{y \in \tilde{S}_i} (y - \bar{\tilde{m}}_i)^2. \tag{95}
\]

Thus, \( (1/(n_1 + n_2)) \tilde{z}_1^2 + \tilde{z}_2^2 \) is an estimate of the variance of the pooled data, and \( \tilde{z}_1^2 + \tilde{z}_2^2 \) is called the total within-class scatter of the projected samples. The Fisher linear discriminant employs that linear function \( w^T x \) for which the criterion function

\[
 J(w) = \frac{|\bar{\tilde{m}}_1 - \bar{\tilde{m}}_2|^2}{\tilde{z}_1^2 + \tilde{z}_2^2} \tag{96}
\]

is maximum (and independent of \( |w| \)). While the \( w \) maximizing \( J() \) leads to the best separation between the two projected sets (in the sense just described), we will also need a threshold criterion before we have a true classifier. We first consider how to find the optimal \( w \), and later turn to the issue of thresholds.

To obtain \( J() \) as an explicit function of \( w \), we define the scatter matrices \( S_1 \) and \( S_w \) by

\[
 S_1 = \sum_{x \in S_1} (x - m_1)(x - m_1)^T \tag{97}
\]

and

\[
 S_w = S_1 + S_2. \tag{98}
\]

Then we can write

\[
 \tilde{z}_1^2 = \sum_{y \in \tilde{S}_1} (w^T x - w^T m_1)^2 = \sum_{x \in \tilde{S}_1} w^T (x - m_1)(x - m_1)^T w = w^T S_w w, \tag{99}
\]

therefore the sum of these scatters can be written

\[
 \tilde{z}_1^2 + \tilde{z}_2^2 = w^T S_w w. \tag{100}
\]

Similarly, the separations of the projected means obeys

\[
 (\bar{\tilde{m}}_1 - \bar{\tilde{m}}_2)^2 = (w^T m_1 - w^T m_2)^2 = w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_w w, \tag{101}
\]

where

\[
 S_w = (m_1 - m_2)(m_1 - m_2)^T. \tag{102}
\]
We call $S_B$ the \textit{within-class scatter matrix}. It is proportional to the sample covariance matrix for the pooled $d$-dimensional data. It is symmetric and positive semidefinite, and it is usually nonsingular if $n > d$. Likewise, $S_S$ is called the \textit{between-class scatter matrix}. It is also symmetric and positive semidefinite, but because it is the outer product of two vectors, its rank is at most one. In particular, for any $w$, $S_B w$ is in the direction of $m_1 - m_2$, and $S_S$ is quite singular.

In terms of $S_B$ and $S_S$, the criterion function $J(\cdot)$ can be written as

$$J(w) = \frac{w' S_B w}{w' S_S w}. \quad (103)$$

This expression is well known in mathematical physics as the generalized Rayleigh quotient. It is easy to show that a vector $w$ that maximizes $J(\cdot)$ must satisfy

$$S_B w = \lambda S_S w,$$  \quad (104)

for some constant $\lambda$, which is a generalized eigenvalue problem (Problem 42). This can also be seen informally by noting that at an extremum of $J(\cdot)$ a small change in $w$ in Eq. 103 should leave unchanged the ratio of the numerator to the denominator. If $S_B$ is nonsingular we can obtain a conventional eigenvalue problem by writing

$$S_B^{-1} S_S w = \lambda w.$$  \quad (105)

In our particular case, it is unnecessary to solve for the eigenvalues and eigenvectors of $S_B^{-1} S_S$ due to the fact that $S_B w$ is always in the direction of $m_1 - m_2$. Because the scale factor for $w$ is immaterial, we can immediately write the solution for the $w$ that optimizes $J(\cdot)$:

$$w = S_B^{-1} (m_1 - m_2). \quad (106)$$

Thus, we have obtained $w$ for Fisher’s linear discriminant—the linear function yielding the maximum ratio of between-class scatter to within-class scatter. (The solution $w$ given by Eq. 106 is sometimes called the canonical variate.) Thus the classification has been converted from a $d$-dimensional problem to a hopefully more manageable one-dimensional one. This mapping is many-to-one, and in theory it cannot possibly reduce the minimum achievable error rate if we have a very large training set. In general, one is willing to sacrifice some of the theoretically attainable performance for the advantages of working in one dimension. All that remains is to find the threshold, that is, the point along the one-dimensional subspace separating the projected points.

When the conditional densities $p(x | c_i)$ are multivariate normal with equal covariance matrices $\Sigma$, we can calculate the threshold directly. In that case we recall from Chapter 2 that the optimal decision boundary has the equation

$$w' x + w_0 = 0 \quad (107)$$

where

$$w = \Sigma^{-1} (\mu_1 - \mu_2),$$ \quad (108)

and where $w_0$ is a constant involving $w$ and the prior probabilities. If we use sample means and the sample covariance matrix to estimate $\mu_i$ and $\Sigma$, we obtain a vector in the same direction as the $w$ of Eq. 108 that maximized $J(\cdot)$. Thus, for the normal, equal-covariance case, the optimal decision rule is merely to decide $c_0$ if Fisher’s linear discriminant exceeds some threshold, and to decide $c_0$ otherwise. More generally, if we smooth the projected data, or fit it with a univariate Gaussian, we then should choose $w_0$ where the posteriors in the one-dimensional distributions are equal.

The computational complexity of finding the optimal $w$ for the Fisher linear discriminant (Eq. 106) is dominated by the calculation of the within-category total scatter and its inverse, an $O(d^2 n)$ calculation.

### 3.8.3 Multiple Discriminant Analysis

For the $c$-class problem, the natural generalization of Fisher’s linear discriminant involves $c - 1$ discriminant functions. Thus, the projection is from a $d$-dimensional space to a $(c - 1)$-dimensional space, and it is tacitly assumed that $d \geq c$. The generalization for the within-class scatter matrix is obvious:

$$S_W = \sum_{i=1}^c S_i \quad (109)$$

where, as before,

$$S_i = \sum_{x \in S_i} (x - m_i)(x - m_i)' \quad (110)$$

and

$$m_i = \frac{1}{n_i} \sum_{x \in S_i} x. \quad (111)$$

The proper generalization for $S_B$ is not quite so obvious. Suppose that we define a total mean vector $m$ and a total scatter matrix $S_T$ by

$$m = \frac{1}{n} \sum_{x} x = \frac{1}{n} \sum_{i=1}^c n_i m_i \quad (112)$$

and

$$S_T = \sum_{x} (x - m)(x - m)'. \quad (113)$$

Then it follows that

$$S_T = \sum_{i=1}^c \sum_{x \in S_i} (x - m_i + m_i - m)(x - m_i + m_i - m)' = \sum_{i=1}^c \sum_{x \in S_i} (x - m_i)(x - m_i)' + \sum_{i=1}^c \sum_{x \in S_i} (m_i - m)(m_i - m)' = S_B + \sum_{i=1}^c n_i (m_i - m)(m_i - m)'. \quad (114)$$