WITHIN-CLASS SCATTER

BETWEEN-CLASS SCATTER

We call $\mathbf{S}_W$ the *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, $\mathbf{S}_B$ is called the *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 $\mathbf{w}$, $\mathbf{S}_B\mathbf{w}$ is in the direction of $\mathbf{m}_1 - \mathbf{m}_2$, and $\mathbf{S}_B$ is quite singular.

In terms of $\mathbf{S}_B$ and $\mathbf{S}_W$, the criterion function $J(\cdot)$ can be written as

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}. \tag{103}$$

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

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}, \tag{104}$$

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

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}. \tag{105}$$

In our particular case, it is unnecessary to solve for the eigenvalues and eigenvectors of $\mathbf{S}_W^{-1}\mathbf{S}_B$ due to the fact that $\mathbf{S}_B\mathbf{w}$ is always in the direction of $\mathbf{m}_1 - \mathbf{m}_2$. Because the scale factor for $\mathbf{w}$ is immaterial, we can immediately write the solution for the $\mathbf{w}$ that optimizes $J(\cdot)$:

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{106}$$

Thus, we have obtained $\mathbf{w}$ for Fisher's linear discriminant—the linear function yielding the maximum ratio of between-class scatter to within-class scatter. (The solution $\mathbf{w}$ 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 $p(\mathbf{x}|\omega_i)$ are multivariate normal with equal covariance matrices $\boldsymbol{\Sigma}$, we can calculate the threshold directly. In that case we recall from Chapter 2 that the optimal decision boundary has the equation

$$\mathbf{w}^t\mathbf{x} + w_0 = 0 \tag{107}$$

where

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{108}$$

and where w_0 is a constant involving $\mathbf{w}$ and the prior probabilities. If we use sample means and the sample covariance matrix to estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$, we obtain a vector in