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

$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i, \tag{89}$$

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

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathbf{m} + \sum_{i=1}^{d'} a_{ki} \mathbf{e}_i \right) - \mathbf{x}_k \right\|^2 \tag{90}$$

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 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 letters O and Q, PCA might discover the gross features that characterize Os and Qs, 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$, n_1 in the subset $\mathcal{D}_1$ labeled ω_1 and n_2 in the subset $\mathcal{D}_2$ labeled ω_2. If we form a linear combination of the components of $\mathbf{x}$, we obtain the scalar dot product

$$y = \mathbf{w}^t \mathbf{x} \tag{91}$$

and a corresponding set of n samples $y_1, \ldots, y_n$ divided into the subsets $\mathcal{Y}_1$ and $\mathcal{Y}_2$. Geometrically, if $\|\mathbf{w}\| = 1$, each y_i is the projection of the corresponding $\mathbf{x}_i$ onto