directly for the eigenvectors $\mathbf{w}_i$. Because $\mathbf{S}_B$ is the sum of c matrices of rank one or less, and because only $c-1$ of these are independent, $\mathbf{S}_B$ is of rank $c-1$ or less. Thus, no more than $c-1$ of the eigenvalues are nonzero, and the desired weight vectors correspond to these nonzero eigenvalues. If the within-class scatter is isotropic, the eigenvectors are merely the eigenvectors of $\mathbf{S}_B$, and the eigenvectors with nonzero eigenvalues span the space spanned by the vectors $\mathbf{m}_i - \mathbf{m}$. In this special case the columns of $\mathbf{W}$ can be found simply by applying the Gram-Schmidt orthonormalization procedure to the $c-1$ vectors $\mathbf{m}_i - \mathbf{m}$, $i = 1, \ldots, c-1$. Finally, we observe that in general the solution for $\mathbf{W}$ is not unique; the allowable transformations include rotating and scaling the axes in various ways. These are all linear transformations from a $(c-1)$-dimensional space to a $(c-1)$-dimensional space, however, and do not change things in any significant way; in particular, they leave the criterion function $J(\mathbf{W})$ invariant and the classifier unchanged.

If we have very little data, we would tend to project to a subspace of low dimension, while if there are more data, we can use a higher dimension, as we shall explore in Chapter 9. Once we have projected the distributions onto the optimal subspace (defined as above), we can use the methods of Chapter 2 to create our full classifier.

As in the two-class case, multiple discriminant analysis primarily provides a reasonable way of reducing the dimensionality of the problem. Parametric or nonparametric techniques that might not have been feasible in the original space may work well in the lower-dimensional space. In particular, it may be possible to estimate separate covariance matrices for each class and use the general multivariate normal assumption after the transformation where this could not be done with the original data. In general, if the transformation causes some unnecessary overlapping of the data and increases the theoretically achievable error rate, then the problem of classifying the data still remains. However, there are other ways to reduce the dimensionality of data, and we shall encounter this subject again in Chapter 10. We note that there are also alternative methods of discriminant analysis—such as the selection of features based on statistical significance—some of which are given in the references for this chapter. Of these, Fisher's method remains a fundamental and widely used technique.

*3.9 EXPECTATION-MAXIMIZATION (EM)

We saw in Chapter 2, Section 2.10 how we could classify a test point even when it has missing features. We can now extend our application of maximum-likelihood techniques to permit the *learning* of parameters governing a distribution from training points, some of which have missing features. If we had uncorrupted data, we could use maximum-likelihood, i.e., find $\hat{\boldsymbol{\theta}}$ that maximized the log-likelihood $l(\boldsymbol{\theta})$. The basic idea in the expectation-maximization or EM algorithm is to iteratively estimate the likelihood given the data that is present. The method has precursors in the Baum-Welch algorithm we will consider in Section 3.10.6.

Consider a full sample $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of points taken from a single distribution. Suppose, though, that here some features are missing; thus any sample point can be written as $\mathbf{x}_k = \{\mathbf{x}_{kg}, \mathbf{x}_{kb}\}$, i.e., comprising the "good" features and the missing, or "bad" ones (Chapter 2, Section 2.10). For notational convenience we separate these individual *features* (not samples) into two sets, $\mathcal{D}_g$ and $\mathcal{D}_b$ with $\mathcal{D} = \mathcal{D}_g \cup \mathcal{D}_b$ being the union of such features.