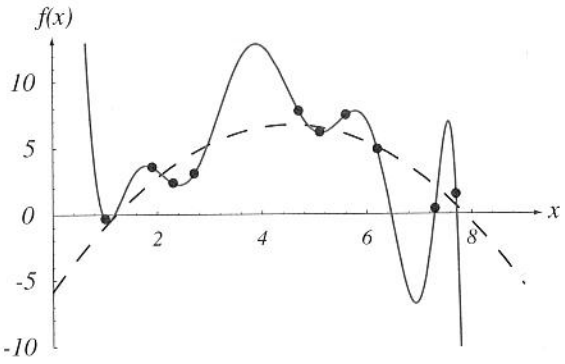

FIGURE 3.4. The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because it would lead to better predictions for new samples.

estimate $\hat{\boldsymbol{\Sigma}}$ accordingly. Such estimation requires proper normalization of the data (Problem 37).

SHRINKAGE

An intermediate approach is to assume a weighted combination of the equal and individual covariances, a technique known as *shrinkage* (also called regularized discriminant analysis), because the individual covariances "shrink" toward a common one. If i is an index on the c categories in question, we have

$$\boldsymbol{\Sigma}_i(\alpha) = \frac{(1-\alpha)n_i\boldsymbol{\Sigma}_i + \alpha n\boldsymbol{\Sigma}}{(1-\alpha)n_i + \alpha n}, \tag{76}$$

for $0 < \alpha < 1$. Additionally, we could "shrink" the estimate of the (assumed) common covariance matrix toward the identity matrix, as

$$\boldsymbol{\Sigma}(\beta) = (1-\beta)\boldsymbol{\Sigma} + \beta\mathbf{I}, \tag{77}$$

for $0 < \beta < 1$ (Computer exercise 8). (Such methods for simplifying classifiers have counterparts in regression, generally known as *ridge regression*.)

Our short, intuitive discussion here will have to suffice until Chapter 9, where we will explore the crucial issue of controlling the complexity or expressive power of a classifier for optimum performance.

*3.8 COMPONENT ANALYSIS AND DISCRIMINANTS

One approach to coping with the problem of excessive dimensionality is to reduce the dimensionality by combining features. Linear combinations are particularly attractive because they are simple to compute and analytically tractable. In effect, linear methods project the high-dimensional data onto a lower dimensional space. There are two classical approaches to finding effective linear transformations. One approach—known as Principal Component Analysis or PCA—seeks a projection that best represents the data in a least-squares sense. Another approach—known as Multiple Discriminant Analysis or MDA—seeks a projection that best separates the data in a least-squares sense. We consider each of these approaches in turn.