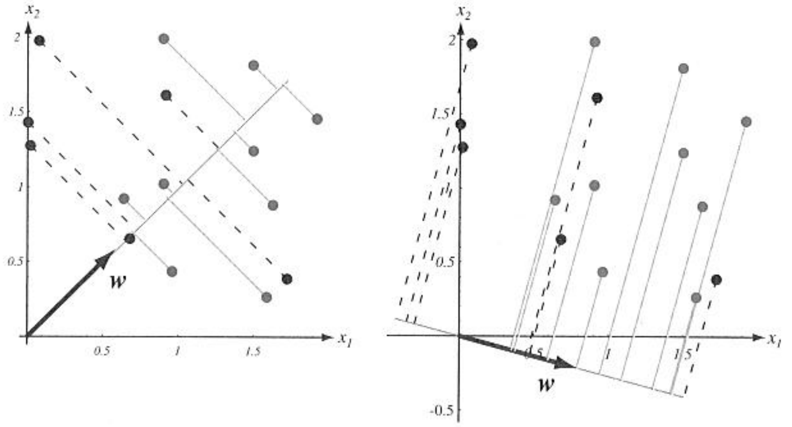

FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points.

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 ω_1 fall more or less into one cluster while those labeled ω_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 $\mathbf{m}_i$ is the d-dimensional sample mean given by

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}, \tag{92}$$

then the sample mean for the projected points is given by

$$\begin{aligned} \tilde{m}_i &= \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y \\ &= \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i \end{aligned} \tag{93}$$

and is simply the projection of $\mathbf{m}_i$.

It follows that the distance between the projected means is

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)| \tag{94}$$

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

---

SCATTER

WITHIN-CLASS SCATTER

SCATTER MATRICES