Website Appendix A
Methodology for the Discriminant Example

Discriminant Problem

Let $\mathbf{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T$ be drawn with probability .5 from either of two groups G_1 : $\mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ or G_2 : $\mathrm{N}\left(\begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$. For a fixed estimation sample size, $l \in \{10, 50, 100, 200, 400\}$ and a given trial, *T*, within this sample size, results are based on the methodology illustrated in *Table WA-1*.

Table WA-1: Discriminant Methodology

Sample Size = l				
Trial = T | l				
		Predicted by Model (L or Q)		
Estimation Sample		G_1	G_2	
l_1=.5*l*	*True Group* G_1	N_{11}	N_{12}	N_1=500
l_2=.5*l*	G_2	N_{21}	N_{22}	N_2=500
$l=l_1+l_2$		$\hat{N}_1$	$\hat{N}_2$	*N=1,000*

On trial *T* (given *l*), parameters of each model (*L* ≡ linear and *Q* ≡ quadratic) are estimated on the basis of *l* observations drawn according to the priors (.5/.5) . Each instantiated model is then used to predict the class membership for each of the (same) *N*=1,000 observations drawn randomly from the two populations, again respecting the priors, or N_g = 500 from each group, *g*=1,2.

The error rate on trial *T* for model *L* or *Q* is the percent of observations misclassified by the model calculated as $R_{emp|T}^{L\ or\ Q} = \frac{N_{12}^{L\,or\,Q} + N_{21}^{L\,or\,Q}}{1{,}000}$. The mean error rate and standard deviation are computed directly from the trial-by-trial results; i.e., the mean is the arithmetic average of these values over $T=100$ trials for a given *l*.

Decision Boundaries

Given the parameters of G_1 and G_2 above, the classification rule that minimizes the expected misclassification loss is given by expression (WA1). (See Johnson and Wichern, 2002.)

$$-\frac{1}{2}\mathbf{x}^T\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)\mathbf{x}+\left(\mu_1^T\,\Sigma_1^{-1}-\mu_2^T\,\Sigma_2^{-1}\right)\mathbf{x}-k\geq\ln\left[\left(\frac{c(1\mid 2)}{c(2\mid 1)}\right)\left(\frac{\pi_2}{\pi_1}\right)\right] \qquad \text{(WA1)}$$

Solving for the constant *k* yields a quadratic equation in $\mathbf{x}=\left(x_1 \quad x_2\right)^T$. Using the parameters for the present problem, we find that *Q* is described by the quadratic equation (WA2).

$$x_2 = 3\ \left(\ -4/3+2/3\,x_1+.57735\left(\ .287682-1.33227\times10^{-15}\,x_1+x_1^2\right)^{1/2}\ \right) \qquad \text{(WA2)}$$

Using the pooled *vcv* matrix with the mean vectors yields the linear boundary (*L*) given by equation (WA3).

$$x_2 = -4+4x_1 \qquad \text{(WA3)}$$

Iso-Probability Contours for Bivariate Normal

Thomasian (1969) provides the expression (WA4) in μ_1, μ_2, σ_1, σ_2, and ρ for the iso-probability contours of a bivariate Normal density.

$$g(x_1, x_2) = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\times\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right] \qquad \text{(WA4)}$$

Setting $g(x_1, x_2) = \delta$ for various $\delta > 0$ determines concentric ellipses in the x_1, x_2 plane.

References

Johnson, Richard A. and Dan W. Wichern (2002), *Applied Multivariate Statistical Analysis* (5th edition), Prentice-Hall, Inc

Thomasian, Aram J. (1969), *The Structure of Probability Theory with Applications*, McGraw-Hill, Inc