Linear and Quadratic Discriminant Analysis

Eric M. Schibli, Matthew J. Courtemanche, and Nicholas R. Lee-Horne

February 20th, 2018

$$P(G=k|X=x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

For Gaussian distribution &

$$f_k(x) = \frac{\pi_k}{\Sigma_k^{1/2}}\exp(-\frac{1}{2\Sigma_k^2}(x-\mu_k)^2)$$

$$\log\frac{P(G=k|X=x)}{P(G=l|X=x)=} = \log\frac{\pi_k}{\pi_l} - \frac{1}{2}\log\frac{\Sigma_k}{\Sigma_l} - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_l)^T\Sigma_k^{-1}(x-\mu_l)$$

$$\delta_k = \log\pi_k - \frac{1}{2}\log\Sigma_k - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)$$

If $\Sigma_k = \Sigma$:

(LDA)

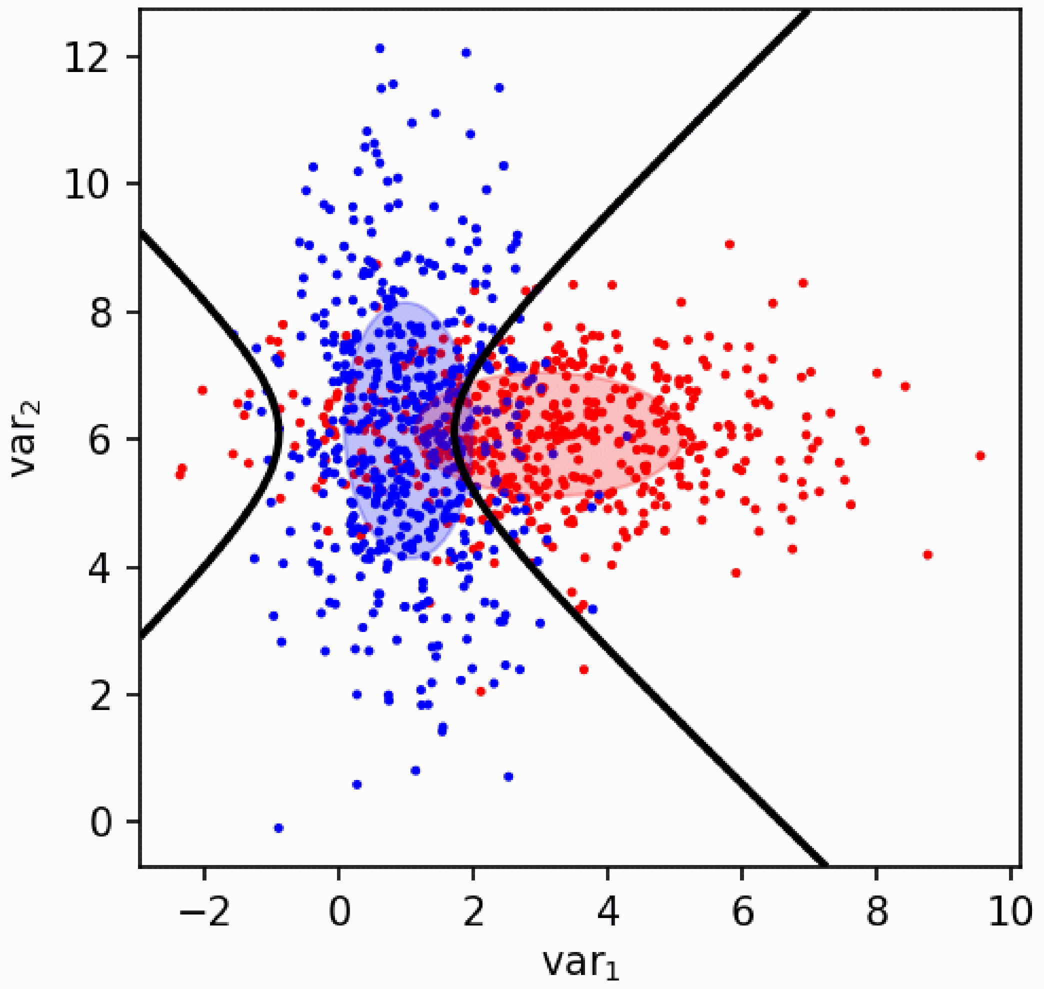
$$\delta_k = \log\pi_k + x^{\mathrm{T}}\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$$

LDA and QDA

- LDA and QDA are **classification** methods
- Assume k classes are normally distributed in **m**-dimensional space.
- Means and covariance matrices taken from training data
- Assigns observation *x* to class with greatest discriminant $\delta(x)$.
- Decision boundaries are contours of equal δ

QDA

- Decision boundaries in QDA are parabolic contours of equal $\delta_k = \delta_k$

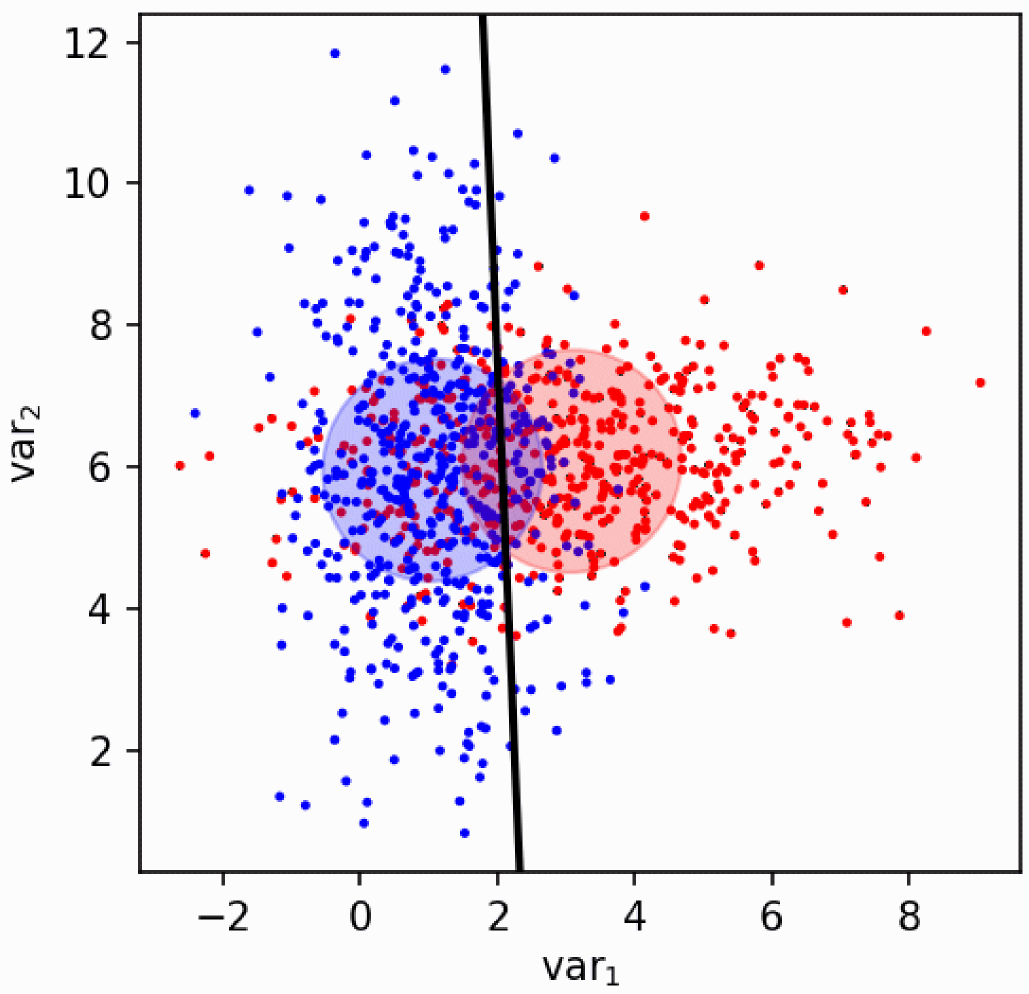


$$\delta_k = \log \pi_k - \frac{1}{2}\log \Sigma_k - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)$$

LDA

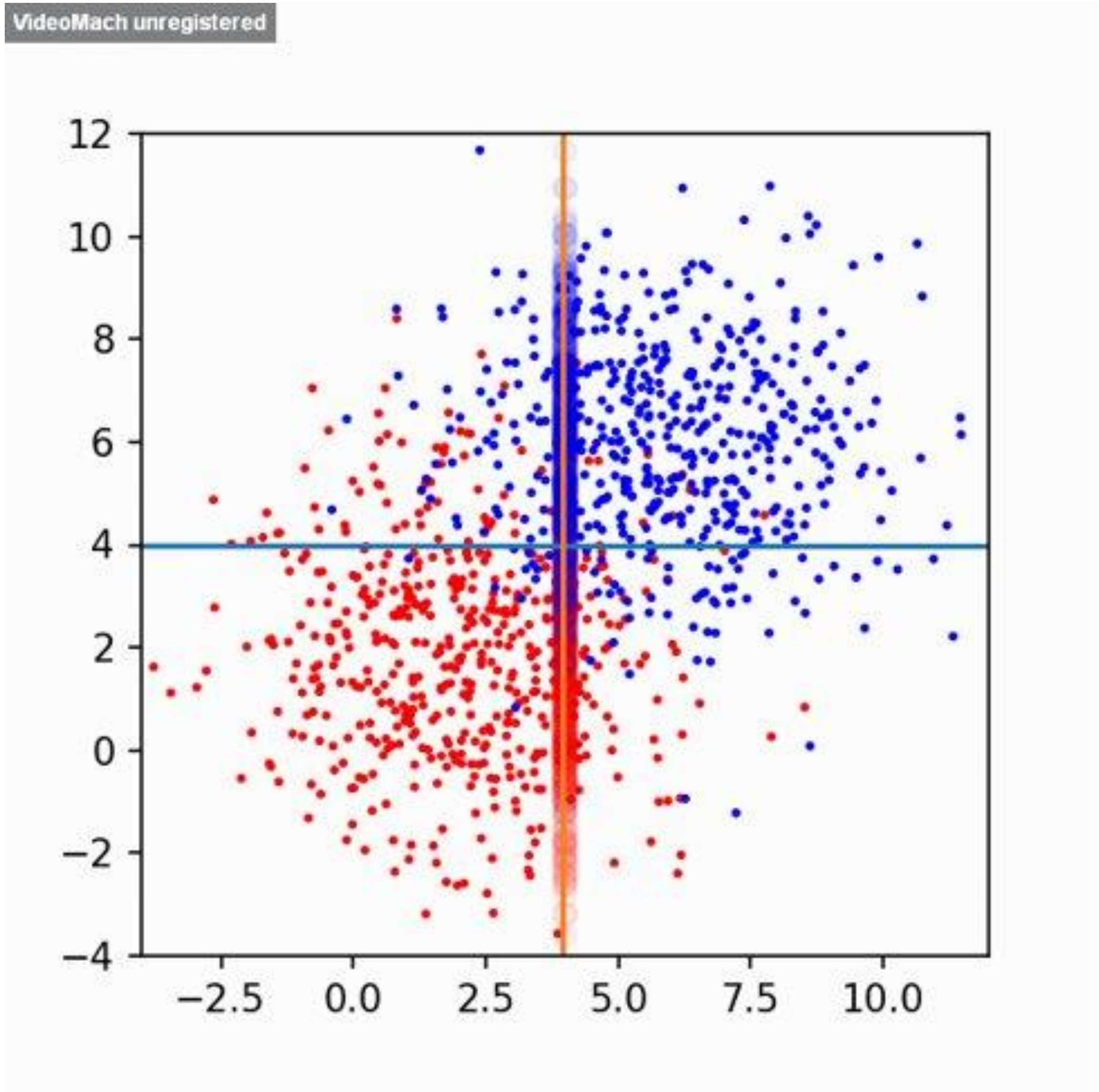
- LDA makes the simplifying assumption that the classes have **identical covariances.**
- This reduces the decision boundaries to lines

$$\delta_k = \log \pi_k + x^{\mathrm{T}}\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$$



LDA

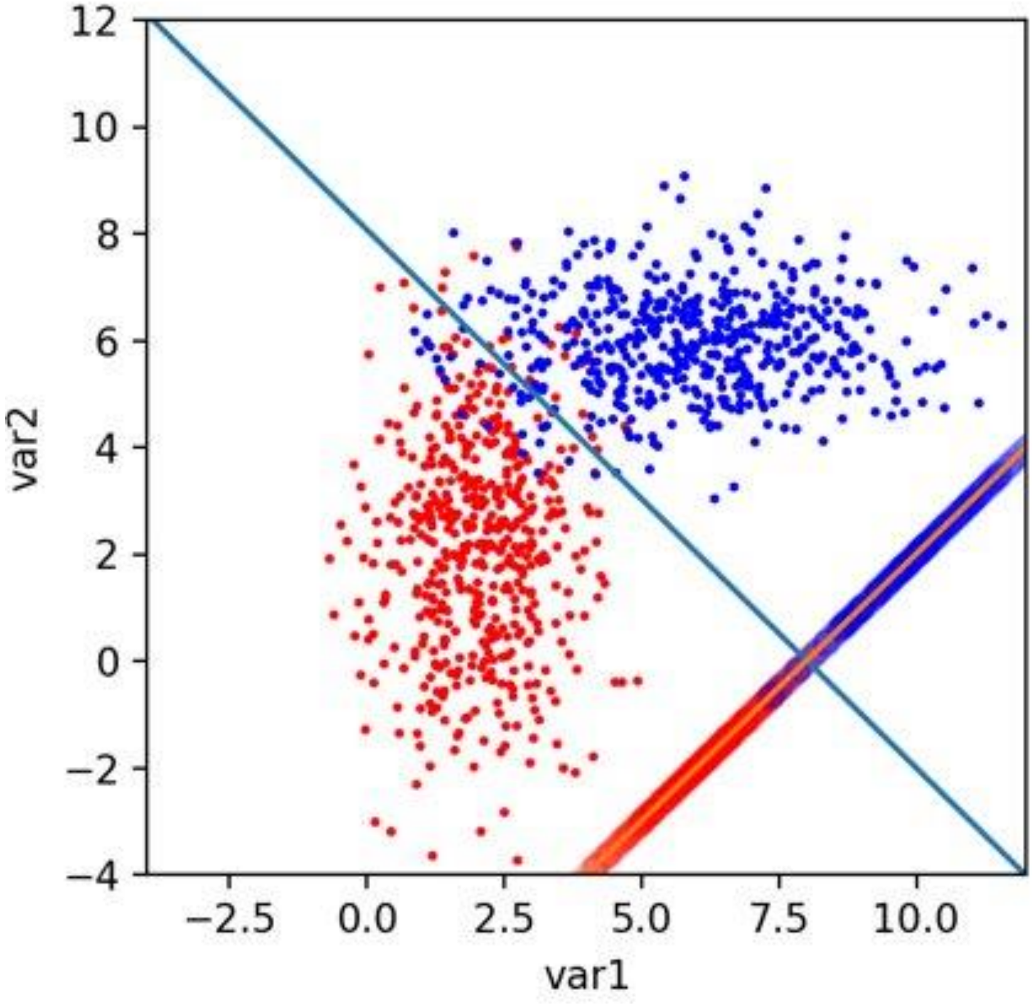
- These decision boundaries produce new axes perpendicular to the boundaries that maximize the separation between the classes



$$\delta_k = \log \pi_k + x^{\mathrm{T}}\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$$

LDA

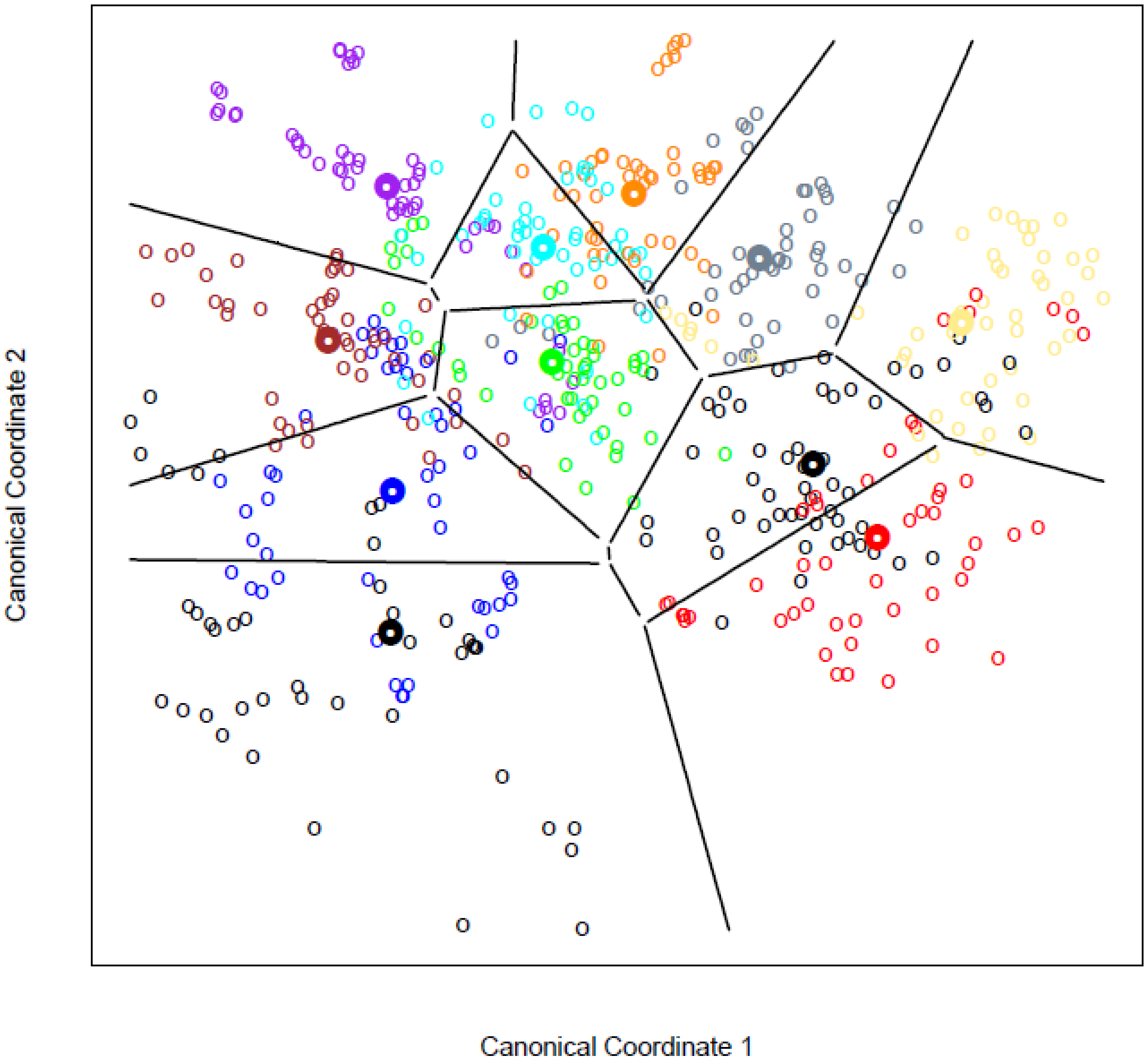
- Projecting the data onto these new axes effectively reduces dimensionality to (p-1)



$$\delta_k = \log \pi_k + x^{\mathrm{T}} \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k$$

Summary

- Both LDA and especially QDA work well on a very wide range of real datasets
- In particular, LDA is often used for dimensionality redution prior to further analysis



Vowel training data, from *Elements of Statistical Learning* (Hastie et al.)

Appendix: Matthews Correlation Coefficient

- Useful metric for highly imbalanced classes
- Essentially the correlation coefficient between the observed and predicted binary classifications
- Ranges from -1 to 1

$$MCC = \frac{(TP \cdot TN) - (TP \cdot TN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TP + FP) \cdot (TP + FN)}}$$

$$MCC = \frac{\frac{TP}{N} - S \cdot P}{\sqrt{P \cdot S \cdot (1 - S) \cdot (1 - P)}}$$

$$S = \frac{TP + FN}{N}$$

$$P = \frac{TP + FP}{N}$$