Fisher’s linear discriminant analysis

Main idea ................................................................. 2
Computation of $a$ ....................................................... 3
Computation of $a$ ....................................................... 4
Alternative method: classification using maximum likelihood ................................. 5
Closing thoughts, possible extensions ........................................... 6
Main idea

- Suppose that we have an $n \times p$ data matrix consisting of $g$ different groups. How can we classify new observations into one of these groups?
- Fisher:
  - Look for the linear function $Xa$ which maximizes the ratio of the between-groups sum of squares to the within-groups sum of squares.
  - Compute average score $(\bar{x}_i)'a$ for each group $i = 1, \ldots, g$.
  - Compute the score $x_{new}'a$ for the new observation.
  - Classify the new observation in group $j$ if $|x_{new}'a - (\bar{x}_j)'a| < |x_{new}'a - (\bar{x}_i)'a|$ for all $i \neq j$.
- See explanation on board.

Computation of $a$

- Consider the total sum of squares of $y = Xa$: $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- This can be split into a between-groups sum of squares and a within-groups sum of squares:
  - The between-groups sum of squares can be written as $a'Ba$, where $B = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$.
  - The within-groups sum of squares can be written as $a'Wa$, where $W = (n_1S_1 + \cdots + n_gS_g)/(n - g)$ is the pooled covariance matrix estimate.
- Find $a$ that maximizes the ratio $(a'Ba)/(a'Wa)$. (Note that $a$ can be re-scaled without affecting this ratio.)

Computation of $a$

- (Th. 11.5.1 of MKB) The vector $a$ that maximizes $(a'Ba)/(a'Wa)$ is given by the eigenvector of $W^{-1}B$ corresponding to the largest eigenvalue. (See proof in handout.)

- In the case of 2 groups, $a$ simplifies: $a = W^{-1}(\bar{x}_1 - \bar{x}_2)$. (Without proof).

- See R-code.
Alternative method: classification using maximum likelihood

- Suppose the exact distributions of the populations $\Pi_1, \ldots, \Pi_g$ are known.
- Then the maximum likelihood discriminant rule for allocating a new observation is to allocate it to the population which gives the largest likelihood to $x$, i.e., to the population with the highest density at the point $x$. See picture on overhead.
- If the exact distributions are unknown, but we know for example that the populations are all multivariate normal, then we can first estimate their parameters, and then use the above rule. This is the sample maximum likelihood discriminant rule.

Closing thoughts, possible extensions

- The way we introduced Fisher’s linear discriminant analysis, no assumptions are needed. It is simply a sensible rule to try to classify observations.
- But if we have two groups from two multivariate normal distributions with the same covariance matrix, then Fisher’s linear discriminant analysis corresponds exactly to the maximum likelihood rule for classification. (For more than two groups this is generally not the case.)
- Possible extensions:
  - We can use the first $k$ eigenvectors of $W^{-1}B$ if we want to summarize the distinction between the groups in $k$-dimensions. This corresponds to projecting the data on a plane in $\mathbb{R}^k$.
  - Quadratic discriminant analysis: boundaries between classes are quadratic rather than linear.