# FINM 331: DATA ANALYSIS FOR FINANCE AND STATISTICS
# FALL 2015
# LINEAR DISCRIMINANT ANALYSIS

## 1. SUPERVISED VS UNSUPERVISED

- supervised learning: data divided into training set and test set; use training set to get parameters in model, apply to predict results for data point in test set
- unsupervised learning: no such division; infer structure from data directly
- examples:
  - unsupervised: PCA, FA, CCA, CA, MDS
  - supervised: LDA, SVM

## 2. SAMPLE LINEAR DISCRIMINANT ANALYSIS

- departure from discussions of PCA, FA, CCA — we do not have a simple 'population LCA' that provides the statistical underpinning for sample LDA
- we will give something that can be considered a 'population LDA for binary classification' in the last section
- training data set is divided into $g$ different *groups* or *classes*

$$C_1, \dots, C_g$$

that we may regard as subset of $\mathbb{R}^p$
- training data is given in the form of a data matrix $X \in \mathbb{R}^{n \times p}$ parititioned into $g$ blocks

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_g \end{bmatrix} \in \mathbb{R}^{n \times p}$$

where the sublock

$$X_i = \begin{bmatrix} \mathbf{x}_{i,1}^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_{i,n_i}^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{n_i \times p}, \quad i = 1, \dots, g,$$

has all row vectors $\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,1} \in C_i$
- we must have

$$n_1 + \cdots + n_g = n$$

- note that for the row vectors of $X$, we know exactly which class it belongs to, we will use this information to build a *classifier*, a vector $\mathbf{q} \in \mathbb{R}^p$ that allows us to determine, for *any* $\mathbf{x} \in \mathbb{R}^p$, which class it belongs to
- as in the case of PCA and CCA we will find our $\mathbf{q}$ by considering linear combinations of the data matrix
- since a linear combination $a_1 x_1 + \cdots + a_p x_p$ is determined by the coefficient vector $\mathbf{a} \in \mathbb{R}^p$, we start from $X\mathbf{a}$

*Date*: November 19, 2015, version 1.2. Comments, bug reports: lekheng@galton.uchicago.edu.

- given the structure of $X$, we can write

$$X\mathbf{a} = \begin{bmatrix} X_1 \\ \vdots \\ X_g \end{bmatrix} \mathbf{a} = \begin{bmatrix} X_1\mathbf{a} \\ \vdots \\ X_g\mathbf{a} \end{bmatrix} =: \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_g \end{bmatrix} =: \mathbf{y} \in \mathbb{R}^n$$

  where $\mathbf{y}_i \in \mathbb{R}^{n_i}$, $i = 1, \ldots, g$
- let's start by looking at sample means

$$\overline{\mathbf{x}} = \frac{1}{n} X^\mathsf{T}\mathbf{1} \in \mathbb{R}^p, \qquad \overline{\mathbf{x}}_i = \frac{1}{n_i} X_i^\mathsf{T}\mathbf{1} \in \mathbb{R}^p$$

  for $i = 1, \ldots, g$ (recall that we use $\mathbf{1} = [1, \ldots, 1]^\mathsf{T}$ to denote a vector of all ones, irrespective of dimension)
- $\overline{\mathbf{x}}_i$, the sample mean of $X_i$, $i = 1, \ldots, g$, are sometimes called *class means*
- note that

$$n\overline{\mathbf{x}} = n_1\overline{\mathbf{x}}_1 + \cdots + n_g\overline{\mathbf{x}}_g \tag{2.1}$$

  since

$$n\overline{\mathbf{x}} = X^\mathsf{T}\mathbf{1} = [X_1^\mathsf{T}, \ldots, X_g^\mathsf{T}] \begin{bmatrix} \mathbf{1} \\ \vdots \\ \mathbf{1} \end{bmatrix} = X_1^\mathsf{T}\mathbf{1} + \cdots + X_g^\mathsf{T}\mathbf{1} = n_1\overline{\mathbf{x}}_1 + \cdots + n_g\overline{\mathbf{x}}_g$$

- also the sample means of the linear combinations

$$\overline{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} = \mathbf{a}^\mathsf{T}\overline{\mathbf{x}} \in \mathbb{R}, \qquad \overline{\mathbf{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \mathbf{a}^\mathsf{T}\overline{\mathbf{x}}_i \in \mathbb{R}$$

  for $i = 1, \ldots, g$, and where $y_{ij}$ is the $j$th coordinate of $\mathbf{y}_i \in \mathbb{R}^{n_i}$
- note also that

$$n\overline{\mathbf{y}} = n_1\overline{\mathbf{y}}_1 + \cdots + n_g\overline{\mathbf{y}}_g$$

- recall the centering matrix in Handout **3**

$$H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T} \in \mathbb{R}^{n \times n}$$

  which has the effect of centering data matrices

$$HX = X - \mathbf{1}\overline{\mathbf{x}}^\mathsf{T} = X_c \in \mathbb{R}^{n \times p}$$

  we shall write $H_i \in \mathbb{R}^{n_i \times n_i}$ for a centering matrix of size $n_i \times n_i$
- next we are going to look at something called *sum-of-squares matrices*, which are essentially covariance matrices
- consider

$$\mathbf{y}^\mathsf{T}H\mathbf{y} = \mathbf{a}^\mathsf{T}X^\mathsf{T}HX\mathbf{a} =: \mathbf{a}^\mathsf{T}T\mathbf{a} \in \mathbb{R}$$

  where the matrix

$$T := X^\mathsf{T}HX = X^\mathsf{T}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}\right)X \in \mathbb{R}^{p \times p}$$

  is called the *total sum-of-squares matrix*, so called because

$$\mathbf{a}^\mathsf{T}T\mathbf{a} = \mathbf{y}^\mathsf{T}H\mathbf{y} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{\mathbf{y}})^2$$

- clearly $T$ is just the sample covriance matrix of $X$ without the usual $1/(n-1)$ in front, i.e.,

$$T = (n-1)S$$

- next consider
$$\sum_{i=1}^{g} \mathbf{y}_i^\mathsf{T} H_i \mathbf{y}_i = \sum_{i=1}^{g} \mathbf{a}^\mathsf{T} X_i^\mathsf{T} H_i X_i \mathbf{a} =: \mathbf{a}^\mathsf{T} W \mathbf{a} \in \mathbb{R}$$
where the matrix
$$W := \sum_{i=1}^{g} X_i^\mathsf{T} H_i X_i \in \mathbb{R}^{p \times p}$$
is called the *within-group sum-of-squares matrix*
- lastly consider
$$\begin{aligned}\sum_{i=1}^{g} (\overline{\mathbf{y}}_i - \overline{\mathbf{y}})^2 &= \sum_{i=1}^{g} n_i [\mathbf{a}^\mathsf{T}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})]^2 = \sum_{i=1}^{g} n_i \mathbf{a}^\mathsf{T}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^\mathsf{T}\mathbf{a} \\ &= \mathbf{a}^\mathsf{T} \left[ \sum_{i=1}^{g} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^\mathsf{T} \right] \mathbf{a} =: \mathbf{a}^\mathsf{T} B \mathbf{a} \in \mathbb{R}\end{aligned}$$
where the matrix
$$B := \sum_{i=1}^{g} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^\mathsf{T} \in \mathbb{R}^{p \times p}$$
is called the *between-group sum-of-squares matrix*
- the relation between $T$, $W$, $B$ is that
$$\mathbf{a}^\mathsf{T} T \mathbf{a} = \mathbf{a}^\mathsf{T} B \mathbf{a} + \mathbf{a}^\mathsf{T} W \mathbf{a},$$
total = between + within
- alternatively, we may use what's known as the *pooled covariance* matrix $S_{\text{pool}}$ in place of the within-group sum-of-squares matrix $W$,
$$S_{\text{pool}} = \frac{1}{n-g} \sum_{i=1}^{g} (n_i - 1) S_i \in \mathbb{R}^{p \times p}$$
where
$$S_i = \frac{1}{n_i - 1} (X_i - \mathbf{1}\overline{\mathbf{x}}_i^\mathsf{T})(X_i - \mathbf{1}\overline{\mathbf{x}}_i^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{p \times p}$$
is the sample covariance matrix of $X_i \in \mathbb{R}^{n_i \times p}$
- for example, for $g = 2$ and 3,
$$S_{\text{pool}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}, \quad S_{\text{pool}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + (n_3 - 1)S_3}{n_1 + n_2 + n_3 - 3},$$
- since
$$W = (n - g) S_{\text{pool}},$$
the results obtained with $S_{\text{pool}}$ in place of $W$ differ only by a constant multiplicative factor $n - g$
- R. A. Fisher's idea for finding the classifier $\mathbf{q} \in \mathbb{R}^p$ is to let it be the $\mathbf{a} \in \mathbb{R}^p$ that maximizes the ratio of *between-group sum-of-squares* and *within-group sum-of-squares*
- the *Fisher linear discriminant function* is thus
$$f(\mathbf{a}) = \frac{\mathbf{a}^\mathsf{T} B \mathbf{a}}{\mathbf{a}^\mathsf{T} W \mathbf{a}}$$
and its maximizer is given by the following theorem

**Theorem 1.** *Suppose $W \in \mathbb{R}^{p\times p}$ is nonsingular. Let $\mathbf{q}_1 \in \mathbb{R}^p$ be the principal eigenvector of the matrix $W^{-1}B$ corresponding to the eigenvalue $\lambda_1 = \lambda_{\max}(W^{-1}B)$. Then*

$$\mathbf{q}_1 = \operatorname*{argmax}_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}^\mathsf{T} B\mathbf{a}}{\mathbf{a}^\mathsf{T} W\mathbf{a}}$$

*and*

$$\lambda_1 = \max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}^\mathsf{T} B\mathbf{a}}{\mathbf{a}^\mathsf{T} W\mathbf{a}}.$$

- this is a result that holds for any matrices $B, W$ so long as $W$ is nonsingular
- the proof follows by observing that

$$\max\left\{\frac{\mathbf{a}^\mathsf{T} B\mathbf{a}}{\mathbf{a}^\mathsf{T} W\mathbf{a}} : \mathbf{a}\neq\mathbf{0}\right\} = \max\left\{\mathbf{a}^\mathsf{T} B\mathbf{a} : \mathbf{a}^\mathsf{T} W\mathbf{a} = 1\right\} = \lambda_{\max}(W^{-1/2}BW^{-1/2}) = \lambda_{\max}(W^{-1}B)$$

  where the last two steps follow from Problem **5** in Homework **3**
- $\lambda_1$ and $\mathbf{q}_1$ may also be regarded as solutions to the following problem

$$B\mathbf{x} = W\mathbf{x}, \quad \mathbf{x}\neq\mathbf{0},$$

  which is known as a *generalized eigenvalue problem* since it reduces to a usual eigenvalue problem when $W = I$
- now how do we use this vector $\mathbf{q}_1 \in \mathbb{R}^p$ to classify new data into groups?
- the *classification rule* is as follows: given a new data point (a point in the test set) $\mathbf{t} \in \mathbb{R}^p$, we find

$$i = \operatorname{argmin}\{\mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_j) : j = 1, \ldots, g\}$$

  and assign $\mathbf{t}$ to the class $C_i$
- what this does is to assign $\mathbf{t}$ to the class whose mean score is nearest to $\mathbf{q}^\mathsf{T}\mathbf{t}$ since our choice of $i$ (assuming unique) implies that

$$|\mathbf{q}^\mathsf{T}\mathbf{t} - \mathbf{q}^\mathsf{T}\overline{\mathbf{x}}_i| < |\mathbf{q}^\mathsf{T}\mathbf{t} - \mathbf{q}^\mathsf{T}\overline{\mathbf{x}}_j| \qquad \text{for all } j \neq i \tag{2.2}$$

- this "nearest rule" is essentially the heuristic behind Fisher's LDA
- as we said earlier we don't have a 'population LDA' to properly motivate this technique but in practice it works well
- but for the special case $g = 2$ we can actually say more about what Fisher's LDA does and even provide a population LDA model

## 3. BINARY CLASSIFICATION

- the special case $g = 2$ when there are two classes is called *binary classification*
- this comes up a lot because a lot of problems may be formulated in a form requiring a YES or NO answer
  - is this the image of a male person?
  - should we buy this asset?
  - is this email a spam?

  even though if you formulate the question in another way, it may have more than two answers (e.g. should we BUY, HOLD, or SELL this asset?)
- let $g = 2$ and $n_1$ and $n_2$ be the number of training data points in $C_1$ and $C_2$ respectively, i.e.,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times p}, \qquad X_1 \in \mathbb{R}^{n_1\times p}, \qquad X_2 \in \mathbb{R}^{n_2\times p}$$

- in this case the expression for $B$ vastly simplifies and as a result there is no need to even compute eigenvectors

$$B = n_1(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}})(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}})^{\mathsf{T}} + n_2(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}})(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}})^{\mathsf{T}} = \frac{n_1 n_2}{n_1 + n_2}\mathbf{d}\mathbf{d}^{\mathsf{T}} \in \mathbb{R}^{p\times p}$$

  where

$$\mathbf{d} = \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 \in \mathbb{R}^p$$

- this follows from (2.1), $(n_1 + n_2)\overline{\mathbf{x}} = n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2$
- we will assume that $\overline{\mathbf{x}}_1 \neq \overline{\mathbf{x}}_2$ since otherwise we cannot hope to separate $C_1$ and $C_2$ with a hyperplane (see below)
- so $\mathbf{d} \neq \mathbf{0}$ and so $B$ is a rank-1 matrix, which means that $W^{-1}B$ is rank-1 since multiplying a matrix by another nonsingular matrix does not change its rank
- so $W^{-1}B$ has exactly one nonzero eigenvalue $\lambda_1 \in \mathbb{R}$ and all the other $p-1$ eigenvalues of $B$ must be zero since

$$\text{number of nonzero eigenvalues} = \text{rank}$$

- now use

$$\text{sum of eigenvalues} = \text{trace}$$

  and we get

$$\lambda_1 + \underbrace{0 + \cdots + 0}_{p-1} = \operatorname{tr}(W^{-1}B) = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\mathbf{d}^{\mathsf{T}}W^{-1}\mathbf{d}$$

  where we have used the easy fact (exercise for you) that

$$\operatorname{tr}(A\mathbf{x}\mathbf{y}^{\mathsf{T}}) = \mathbf{x}^{\mathsf{T}}A\mathbf{y} = \mathbf{y}^{\mathsf{T}}A\mathbf{x}$$

- hence we duduce that

$$\lambda_1 = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\mathbf{d}^{\mathsf{T}}W^{-1}\mathbf{d}$$

  which must be the largest eigenvalue since all the others are zeroes
- the eigenvector corresponding to $\lambda_1$ is given by

$$\mathbf{q}_1 = W^{-1}\mathbf{d}$$

- this can be easily veified as follows

$$(W^{-1}B)(W^{-1}\mathbf{d}) = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(W^{-1}\mathbf{d}\mathbf{d}^{\mathsf{T}})W^{-1}\mathbf{d} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)W^{-1}\mathbf{d}(\mathbf{d}^{\mathsf{T}}W^{-1}\mathbf{d})$$

  by the associativity of matrix multiplication and since $\mathbf{d}^{\mathsf{T}}W^{-1}\mathbf{d}$ is just a scalar, we may bring it to the front and get

$$= \left(\frac{n_1 n_2}{n_1 + n_2}\mathbf{d}^{\mathsf{T}}W^{-1}\mathbf{d}\right)W^{-1}\mathbf{d} = \lambda_1 W^{-1}\mathbf{d}$$

  as required
- now what is the classification rule?
- the important point to note is that given any vector $\mathbf{a} \in \mathbb{R}^p$ and a value $c \in \mathbb{R}$, we get a *hyperplane*, an affine subspace of dimension $p-1$ in $\mathbb{R}^p$

$$H = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{a}^{\mathsf{T}}\mathbf{x} = c\}$$

- this hyperplane partitions $\mathbb{R}^p$ into two halves,

$$H_+ = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{a}^{\mathsf{T}}\mathbf{x} > c\}, \qquad H_+ = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{a}^{\mathsf{T}}\mathbf{x} < c\}$$

  a point in the test set $\mathbf{t} \in \mathbb{R}^p$ is classified into one of the two classes depending on which half it falls into

- when we had $g > 2$ classes, we had to use the 'nearest rule' in (2.2) but here we just need to find an $\mathbf{a} \in \mathbb{R}^p$ and a $c \in \mathbb{R}$ and use the sign

$$\operatorname{sgn}(\mathbf{a}^\mathsf{T}\mathbf{x} - c)$$

to classify a given point in the test set $\mathbf{t} \in \mathbb{R}^p$

$$\begin{aligned} \mathbf{t} \text{ is assigned to } C_1 &\iff \mathbf{t} \in H_+ \iff \operatorname{sgn}(\mathbf{a}^\mathsf{T}\mathbf{x} - c) = +1, \\ \mathbf{t} \text{ is assigned to } C_2 &\iff \mathbf{t} \in H_- \iff \operatorname{sgn}(\mathbf{a}^\mathsf{T}\mathbf{x} - c) = -1, \end{aligned}$$

note that this only works when there are two classes
- for the Fisher linear classifier $\mathbf{q}_1 \in \mathbb{R}^p$, the hyperplane is called a *discriminant hyperplane*
- given a point in the test set $\mathbf{t} \in \mathbb{R}^p$, we compare

$$\mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_1) \qquad \text{and} \qquad \mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_2)$$

- note that[1]

$$\begin{aligned} \mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_1) > \mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_2) &\iff \mathbf{q}_1^\mathsf{T}\left(\mathbf{t} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) > 0, \\ \mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_1) < \mathbf{q}_1^\mathsf{T}(\mathbf{t} - \overline{\mathbf{x}}_2) &\iff \mathbf{q}_1^\mathsf{T}\left(\mathbf{t} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) < 0, \end{aligned}$$

and so

$$\mathbf{a} = \mathbf{q}_1 \qquad \text{and} \qquad c = \mathbf{q}_1^\mathsf{T}\left(\frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right)$$

- using the expression for $\mathbf{q}_1$, we get the famous Fisher rule for binary classification

$$\begin{aligned} \mathbf{t} \text{ is assigned to } C_1 &\iff (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T} W^{-1}\left(\mathbf{t} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) > 0, \\ \mathbf{t} \text{ is assigned to } C_2 &\iff (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T} W^{-1}\left(\mathbf{t} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) < 0, \end{aligned}$$

where

$$W = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

- if $n_1 = n_2$, then $W$ is a constant positive multiple of $S_1 + S_2$ and since the sign of $(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T} W^{-1}(\mathbf{t} - (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)/2)$ does not depend on scaling by positive scalars, the classification rule may be restated as

$$\begin{aligned} \mathbf{t} \text{ is assigned to } C_1 &\iff (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T} (S_1 + S_2)^{-1}\left(\mathbf{t} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) > 0, \\ \mathbf{t} \text{ is assigned to } C_2 &\iff (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\mathsf{T} (S_1 + S_2)^{-1}\left(\mathbf{t} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) < 0, \end{aligned}$$

bear this in mind for the next section
- the case "$= 0$" occurs with probability zero and we shall not be concerned with this

[1] The point is to relate the binary classification case with the general ($g > 2$) case — on the left is the classifier we used in the general case (for the nearest rule) and on the right is the one we used in the binary case (to define a separating hyperplane).

## 4. POPULATION LDA FOR BINARY CLASSIFICATION

- we assume the two classes $C_1$ and $C_2$ are from two different populations with population mean and covariance given by $\boldsymbol{\mu}_1 \in \mathbb{R}^p$, $\Sigma_1 \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\mu}_2 \in \mathbb{R}^p$, $\Sigma_2 \in \mathbb{R}^{p\times p}$ respectively
- furthermore

$$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \qquad \Sigma_1 \neq \Sigma_2$$

- the population linear discriminant function is

$$f(\mathbf{a}) = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\mathbf{a}^\mathsf{T}\boldsymbol{\mu}_1 - \mathbf{a}^\mathsf{T}\boldsymbol{\mu}_2)^2}{\mathbf{a}^\mathsf{T}\Sigma_1\mathbf{a} - \mathbf{a}^\mathsf{T}\Sigma_2\mathbf{a}} = \frac{[\mathbf{a}^\mathsf{T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\mathbf{a}^\mathsf{T}(\Sigma_1 - \Sigma_2)\mathbf{a}}$$

- as in the previous section, the maximum of $f$ is

$$\mathbf{q}_1 = \operatorname*{argmax}_{\mathbf{a}\neq\mathbf{0}} f(\mathbf{a}) = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \in \mathbb{R}^p$$

  where

$$\Sigma = \Sigma_1 + \Sigma_2 \in \mathbb{R}^{p\times p}$$

  is assumed to be nonsingular
- the threshold $c \in \mathbb{R}$ is

$$c = \mathbf{q}_1^\mathsf{T}\left(\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) = \frac{1}{2}\boldsymbol{\mu}_1^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2^\mathsf{T}\Sigma^{-1}\boldsymbol{\mu}_2$$

- the classification rule is then

$$\begin{aligned} \mathbf{t} \text{ is assigned to } C_1 \quad &\Longleftrightarrow \quad \mathbf{q}_1^\mathsf{T}\mathbf{t} > c, \\ \mathbf{t} \text{ is assigned to } C_2 \quad &\Longleftrightarrow \quad \mathbf{q}_1^\mathsf{T}\mathbf{t} < c \end{aligned}$$

## 5. EXAMPLE: SAMPLE LDA

- this is the Swiss bank notes data set we saw in Slides **2**
- we sill use the following subscripts for convenience

$$g = \text{genuine}, \qquad c = \text{counterfeit}$$

- there are $n_g + n_c = 200$ bank notes with $n_g = 100$ genuine Swiss banknotes and $n_c = 100$ counterfeit ones
- we take $p = 6$ measurements for each bank note:
  - length of bill
  - width of left edge
  - width of right edge
  - bottom margin width
  - top margin width
  - length of image diagonal
- the data matrix is

$$X = \begin{bmatrix} X_g \\ X_c \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \vdots \\ \mathbf{x}_{200}^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{200\times 6}, \quad X_g = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \vdots \\ \mathbf{x}_{100}^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{100\times 6}, \quad X_c = \begin{bmatrix} \mathbf{x}_{101}^\mathsf{T} \\ \vdots \\ \mathbf{x}_{200}^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{100\times 6}$$

- the *between-group sum-of-squares* is

$$\mathbf{a}^\mathsf{T}B\mathbf{a} = 100[(\overline{\mathbf{y}}_g - \overline{\mathbf{y}})^2 + (\overline{\mathbf{y}}_c - \overline{\mathbf{y}})^2]$$

- the *within-group sum-of-squares* is

$$\mathbf{a}^{\mathsf{T}} W \mathbf{a} = \sum_{i=1}^{100} [(\mathbf{y}_g)_i - \overline{\mathbf{y}}_g]^2 + \sum_{i=1}^{100} [(\mathbf{y}_c)_i - \overline{\mathbf{y}}_c]^2$$

  where

$$(\mathbf{y}_g)_i = \mathbf{a}^{\mathsf{T}} \mathbf{x}_i, \qquad (\mathbf{y}_c)_i = \mathbf{a}^{\mathsf{T}} \mathbf{x}_{i+100}, \qquad i = 1, \dots, 100$$

- the between-group sum-of-squares matrix and within-group sum-of-squares matrix are

$$B = 100[(\overline{\mathbf{x}}_g - \overline{\mathbf{x}})(\overline{\mathbf{x}}_g - \overline{\mathbf{x}})^{\mathsf{T}} + (\overline{\mathbf{x}}_c - \overline{\mathbf{x}})(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})^{\mathsf{T}}] \in \mathbb{R}^{6\times 6},$$
$$W = 100(S_g + S_c) \in \mathbb{R}^{6\times 6},$$

  where $S_g$ and $S_c$ are the sample covariance matrices of $X_g$ and $X_c$
- in this case, because $n_g = n_c$,

$$\overline{\mathbf{y}} = \frac{1}{2}(\overline{\mathbf{y}}_g + \overline{\mathbf{y}}_c) \in \mathbb{R}, \qquad \overline{\mathbf{x}} = \frac{1}{2}(\overline{\mathbf{x}}_g + \overline{\mathbf{x}}_c) \in \mathbb{R}^6$$

- the linear classifier can be computed to be

$$\mathbf{q}_1 = W^{-1}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_c) = \begin{bmatrix} 0.000 \\ 0.029 \\ -0.029 \\ -0.039 \\ -0.041 \\ 0.054 \end{bmatrix}$$

- given a new bank note with measurements $\mathbf{t} \in \mathbb{R}^6$, the classification rule is

$$\mathbf{t} \text{ is genuine if } \mathbf{q}_1^{\mathsf{T}}(\mathbf{t} - \overline{\mathbf{x}}) > 0,$$
$$\mathbf{t} \text{ is counterfeit if } \mathbf{q}_1^{\mathsf{T}}(\mathbf{t} - \overline{\mathbf{x}}) < 0$$

- if we apply this to $\mathbf{t}$ in the training set, i.e., $\mathbf{t} \in \{\mathbf{x}_1, \dots, \mathbf{x}_{200}\}$, we will see that one genuine banknote is misclassfied as counterfeit, every other genuine banknote is correctly classified, and all counterfeit banknotes are correctly classified
- the *apparent error rate*, defined to be the fraction of observations in the training set that are misclassified by the classification rule, is therefore

$$\frac{1}{200} = 0.005$$

  or 0.5% which is very good
- the real test, is however to decide whether a banknote $\mathbf{t} \notin \{\mathbf{x}_1, \dots, \mathbf{x}_{200}\}$, one we have never seen before, is genuine or counterfeit

## 6. using lda like pca and cca

- one may also do scatter plots with LDA like those in PCA and CCA
- so far we have only used the principal eigenvector $\mathbf{q}_1$ of the matrix $W^{-1}B \in \mathbb{R}^{p\times p}$ for classification purposes but we may find all the eigenvectors of $W^{-1}B$
- let the EVD of $W^{-1}B$ be

$$W^{-1}B = Q\Lambda Q^{\mathsf{T}}$$

  where $Q = [\mathbf{q}_1, \dots, \mathbf{q}_p]$ are the eigenvectors corresponding to eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$
- we can project the training set $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$, i.e., the rows of the data matrix $X$, onto $\operatorname{span}\{\mathbf{q}_i, \mathbf{q}_j\}$ corresponding to large eigenvalues $\lambda_i, \lambda_j$ and see if the plot shows us anything
- the points from different classes are indicated using different colors or different symbols

- this idea was originally proposed by C. R. Rao, who used the *sample covariance matrix of class means*,
$$S_b = \frac{1}{g}\sum_{i=1}^{g}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^\mathsf{T} \in \mathbb{R}^{p\times p}$$
and the usual sample covariance matrix of $X$,
$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T} \in \mathbb{R}^{p\times p}$$
in place of $B$ and $W$
- in other words, C. R. Rao considered the generalized eigenvectors of
$$S\mathbf{x} = S_b\mathbf{x}, \qquad \mathbf{x} \neq \mathbf{0},$$
or equivalently the stationary points of the optimization problem
$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\mathsf{T} S_b \mathbf{x}}{\mathbf{x}^\mathsf{T} S \mathbf{x}},$$
or equivalently the eigenvectors of
$$S_b^{-1}S \in \mathbb{R}^{p\times p}$$
- what he showed was that if there are $g$ classes, then the first $g-1$ eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_{g-1}$ corresponding to the $g-1$ largest eigenvalues of $S_b^{-1}S$ span a subspace that contains most of the variability between features