Exploratory analysis
Linear discriminant analysis

Olivier Roustant

April 23, 2020

Position of the slides / textbook

Linear Discriminant Analysis (PCA) can be viewed either:

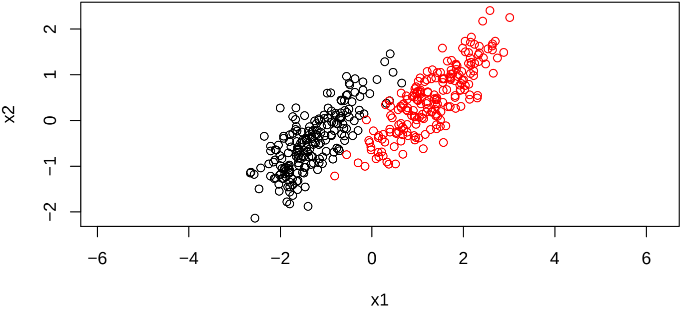
- As a **technique to discover classes in data** (**Fisher's analysis**)
- As a **probabilistic linear method for classification** (prediction)

These slides presents these two facets.

Linear Discriminant Analysis (LDA): Outline

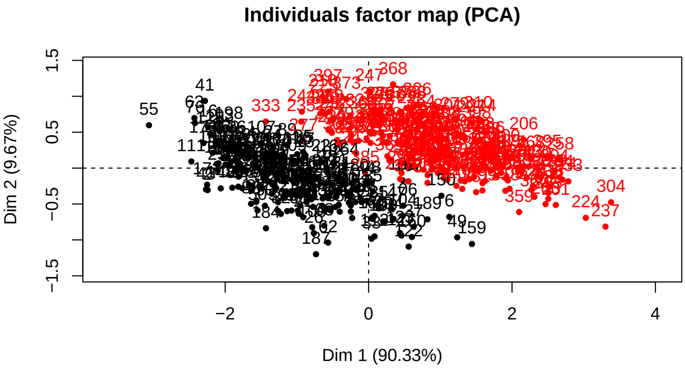
1. **Figures only!**
2. **LDA as an exploratory tool: Theory**
3. **LDA as a classification tool: Theory**

LDA, as an exploratory tool



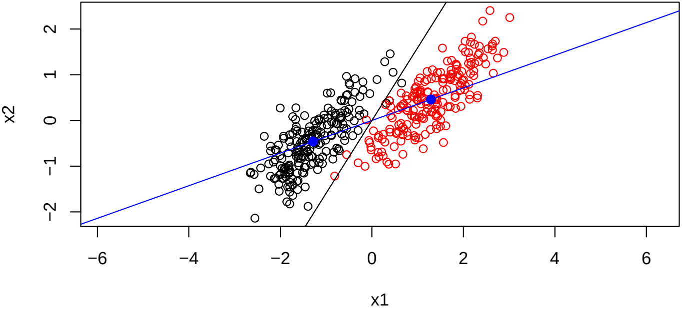
This is a cloud of points, with two classes, in dimension 2 (higher in general).
Can you find two 1D axis ‘suitable’ to identify classes?

LDA, as an exploratory tool



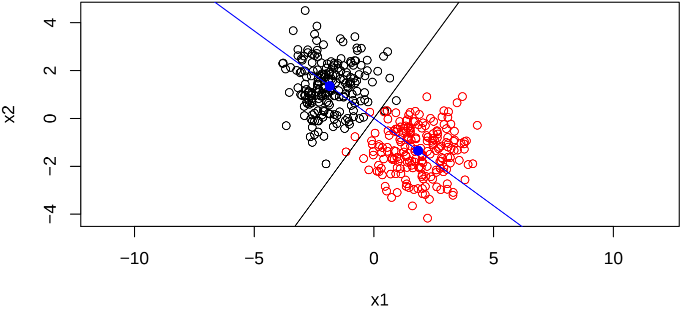
Result of the PCA analysis. Can we do better?

LDA, as an exploratory tool



Result of the LDA analysis. Actually a PCA for the centroids: two data only!
The two axes are orthogonal... for a specific ('Mahalanobis') metric!

LDA, as an exploratory tool

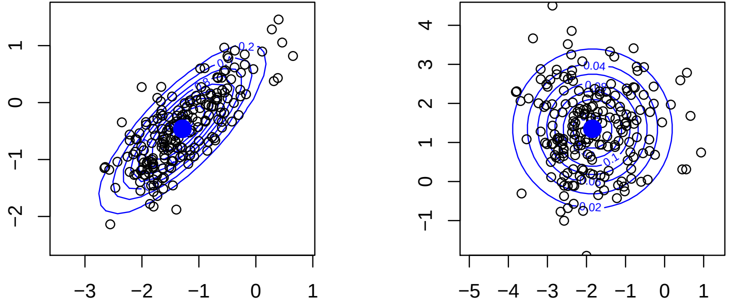


Result of the LDA analysis: visualization for tranformed data.
The two axes are orthogonal for the usual metric.

Mahalanobis metric and 'sphered' data

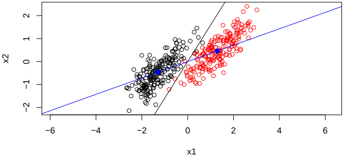
The Mahalanobis metric is such that the covariance matrix is identity.
This is equivalent to **(matricially) reduce or 'sphere'** the data:

$$\mathbf{x} \mapsto \mathrm{Cov}^{-1/2}\mathbf{x}$$

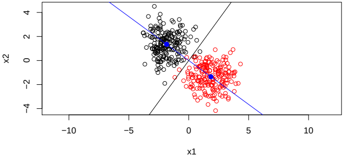


Left: Original data. Right: Reduced data. Level sets are for the multinormal distribution with corresponding covariance matrix.

LDA, as a classification tool



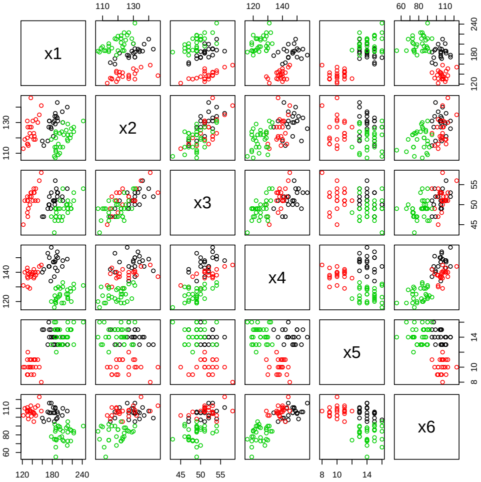
$\Rightarrow$



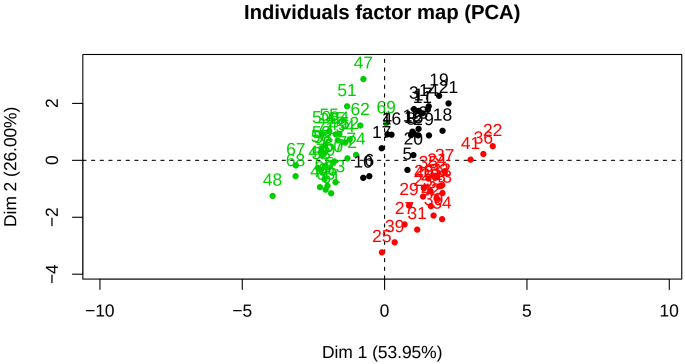
Case of equal group sizes: use sphered data (right) and predict by the class of the nearest centroid (here defined by the line segment bisector).
N.B. This is not optimal when groups have different sizes.

A six dimensional example

Similarly to Fisher's iris data (see notebook), we consider the Lubitsch data for insects. There are 74 data, 6 variables, and 3 classes.

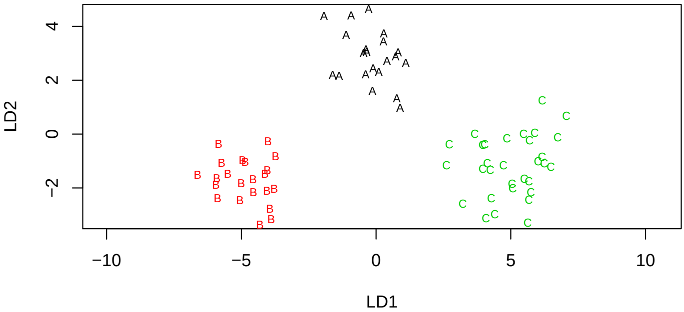


A six dimensional example



Insect dataset. Result of the PCA analysis.

A six dimensional example



Insect dataset. Result of the LDA analysis.

LDA as an exploratory tool: Theory

Notations and assumption

- **X**: a matrix of size $n \times p$, representing the data, partitioned in m classes $\Omega_1, \dots, \Omega_m$ of size $n_1, \dots, n_m$:

	$\mathbf{x}^1$	$\dots$	$\mathbf{x}^j$	$\dots$	$\mathbf{x}^p$	Class
$\mathbf{x}_1$	x_1^1	$\dots$	x_1^j	$\dots$	x_1^p	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{x}_{n_1}$	$x_{n_1}^1$	$\dots$	$x_{n_1}^j$	$\dots$	$x_{n_1}^p$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{x}_{n-n_m+1}$	$x_{n-n_m+1}^1$	$\dots$	$x_{n-n_m+1}^j$	$\dots$	$x_{n-n_m+1}^p$	m
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{x}_n$	x_n^1	$\dots$	x_n^j	$\dots$	x_n^p	m

Notations and assumption

- **G**: a matrix of size $m \times p$, containing the centroids (centor of gravity) of each class: $\mathbf{g}_\ell = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \quad (\ell = 1, \dots, m)$

	$\mathbf{x}^1$	$\dots$	$\mathbf{x}^j$	$\dots$	$\mathbf{x}^p$	Class
$\mathbf{g}_1$	g_1^1	$\dots$	g_1^j	$\dots$	g_1^p	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{g}_m$	g_m^1	$\dots$	g_m^j	$\dots$	g_m^p	m

Notations and assumption

- **G**: a matrix of size $m \times p$, containing the centroids (centor of gravity) of each class: $\mathbf{g}_\ell = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \quad (\ell = 1, \dots, m)$

	$\mathbf{x}^1$	$\dots$	$\mathbf{x}^j$	$\dots$	$\mathbf{x}^p$	Class
$\mathbf{g}_1$	g_1^1	$\dots$	g_1^j	$\dots$	g_1^p	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{g}_m$	g_m^1	$\dots$	g_m^j	$\dots$	g_m^p	m

- Notice that the average of the centroids, weighted by class sizes, coincides with the centroid **g** of the whole dataset:

$$\sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{g}_\ell = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \left(\frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{g}$$

Notations and assumption

- **G**: a matrix of size $m \times p$, containing the centroids (centor of gravity) of each class: $\mathbf{g}_\ell = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \quad (\ell = 1, \dots, m)$

	$\mathbf{x}^1$	$\dots$	$\mathbf{x}^j$	$\dots$	$\mathbf{x}^p$	Class
$\mathbf{g}_1$	g_1^1	$\dots$	g_1^j	$\dots$	g_1^p	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\mathbf{g}_m$	g_m^1	$\dots$	g_m^j	$\dots$	g_m^p	m

- Notice that the average of the centroids, weighted by class sizes, coincides with the centroid **g** of the whole dataset:

$$\sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{g}_\ell = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \left(\frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mathbf{g}$$

- **We assume that $\mathbf{g} = \mathbf{0}$**, i.e. the data have been centered.

Notations and assumption

- **B**: **'between-class'** covariance matrix. It is the covariance matrix of the centroids, weighted by class sizes.

$$\mathbf{B} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{g}_\ell \mathbf{g}_\ell^\top$$

Notations and assumption

- **B**: **'between-class'** covariance matrix. It is the covariance matrix of the centroids, weighted by class sizes.

$$\mathbf{B} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{g}_\ell \mathbf{g}_\ell^\top$$

- **W** : **'within-class'** covariance matrix. It is the covariance matrix of departures to centroids.

$$\mathbf{W} = \frac{1}{n} \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top$$

Notations and assumption

- **B**: **'between-class'** covariance matrix. It is the covariance matrix of the centroids, weighted by class sizes.

$$\mathbf{B} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{g}_\ell \mathbf{g}_\ell^\top$$

- **W** : **'within-class'** covariance matrix. It is the covariance matrix of departures to centroids.

$$\mathbf{W} = \frac{1}{n} \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top$$

Notice that $\mathbf{W} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \left(\frac{1}{n_\ell} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top \right)$ is the (weighted) average of the covariance matrices in each class.
The same within-class covariance matrix is used for all classes $\rightarrow$ (group) homoscedasticity assumption.

Variance decomposition for classes

Property (variance decomposition)

Let $S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ be the covariance matrix of the data. Then,

$$\mathbf{S} = \mathbf{B} + \mathbf{W}$$

[1] This is similar to the formula $\mathbb{E}(Z^2) = \mathrm{Var}(Z) + \mathbb{E}(Z)^2$. To prove it, expand the left hand site by writing $\mathbf{x}_i = (\mathbf{x}_i - \mathbf{g}_\ell) + \mathbf{g}_\ell$, and remark that $\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell) = 0$.

Variance decomposition for classes

Property (variance decomposition)

Let $S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$ be the covariance matrix of the data. Then,

$$\mathbf{S} = \mathbf{B} + \mathbf{W}$$

Proof. Consider one class $\ell \in \{1, \dots, m\}$. Then, we have[1]:

$$\frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top + \mathbf{g}_\ell \mathbf{g}_\ell^\top \qquad (1)$$

[1] This is similar to the formula $\mathbb{E}(Z^2) = \mathrm{Var}(Z) + \mathbb{E}(Z)^2$. To prove it, expand the left hand site by writing $\mathbf{x}_i = (\mathbf{x}_i - \mathbf{g}_\ell) + \mathbf{g}_\ell$, and remark that $\sum_{i \in \Omega_\ell} (\mathbf{x}_i - \mathbf{g}_\ell) = 0$.

Variance decomposition for classes

Property (variance decomposition)

Let $S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$ be the covariance matrix of the data. Then,

$$\mathbf{S} = \mathbf{B} + \mathbf{W}$$

Proof. Consider one class $\ell \in \{1, \dots, m\}$. Then, we have[1]:

$$\frac{1}{n_\ell}\sum_{i\in\Omega_\ell} \mathbf{x}_i\mathbf{x}_i^\top = \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top + \mathbf{g}_\ell\mathbf{g}_\ell^\top \tag{1}$$

Now, multiplying (1) by $\frac{n_\ell}{n}$ and summing w.r.t. ℓ gives: $\mathbf{S} = \mathbf{W} + \mathbf{B}$.

[1] This is similar to the formula $\mathbb{E}(Z^2) = \mathrm{Var}(Z) + \mathbb{E}(Z)^2$. To prove it, expand the left hand site by writing $\mathbf{x}_i = (\mathbf{x}_i - \mathbf{g}_\ell) + \mathbf{g}_\ell$, and remark that $\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell) = 0$.

Problem formulation

The problem (Fisher's approach)

Find a linear combination $\mathbf{a}_1^\top \mathbf{X}$ *maximizing the between-class variance relatively to the within-class variance*:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^\top \mathbf{B} \mathbf{a}}{\mathbf{a}^\top \mathbf{W} \mathbf{a}}$$

Once $\mathbf{a}_1$ found, find $\mathbf{a}_2$, $\mathbf{W}$*-orthogonal to* $\mathbf{a}_1$, maximizing that ratio.
Once $\mathbf{a}_2$ found, find $\mathbf{a}_3$, $\mathbf{W}$*-orthogonal to* $\mathbf{a}_1, \mathbf{a}_2$, maximizing the ratio.
...

N.B. We recall that $\mathbf{a}$ and $\mathbf{b}$ are $\mathbf{W}$-orthogonal if $\mathbf{a}^\top \mathbf{W} \mathbf{b} = 0$.

Main result

Theorem (LDA solution)

The solution of LDA is obtained in two steps:

- Sphere the data with Mahalanobis metric: $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$
- Do PCA on the (sphered) centroids $\mathbf{W}^{-1/2}\mathbf{g}_1, \dots, \mathbf{W}^{-1/2}\mathbf{g}_m$ $\to$ eigenvectors $\mathbf{a}_1^*, \dots, \mathbf{a}_m^*$

The new variables $\mathbf{X}\mathbf{W}^{-1/2}\mathbf{a}_\ell^*$ are called *discriminant variables*.
The $\mathbf{a}_\ell = \mathbf{W}^{-1/2}\mathbf{a}_\ell^*$ are the *discriminant coordinates*.

Main result (proof)

- First observe that when $\mathbf{W} = \mathbf{I}_p$, then LDA = PCA on the centroids, weighted by class sizes.

Main result (proof)

- First observe that when $\mathbf{W} = \mathbf{I}_p$, then LDA = PCA on the centroids, weighted by class sizes.
  Indeed, the numerator of the criterion (Rayleigh ratio) is equal to the variance (inertia) of the projections $\mathbf{a}^\top \mathbf{g}_\ell$ with weights $\frac{n_\ell}{n}$:

$$\mathbf{a}^\top \mathbf{B} \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{a}^\top \mathbf{g}_\ell \mathbf{g}_\ell^\top \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} (\mathbf{a}^\top \mathbf{g}_\ell)^2.$$

Main result (proof)

- First observe that when $\mathbf{W} = \mathbf{I}_p$, then LDA = PCA on the centroids, weighted by class sizes.
  Indeed, the numerator of the criterion (Rayleigh ratio) is equal to the variance (inertia) of the projections $\mathbf{a}^\top \mathbf{g}_\ell$ with weights $\frac{n_\ell}{n}$:

  $$\mathbf{a}^\top \mathbf{B} \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{a}^\top \mathbf{g}_\ell \mathbf{g}_\ell^\top \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} (\mathbf{a}^\top \mathbf{g}_\ell)^2.$$

  The denominator is $\mathbf{a}^\top \mathbf{a}$ is the squared norm of $\mathbf{a}$.

Main result (proof)

- First observe that when $\mathbf{W} = \mathbf{I}_p$, then LDA = PCA on the centroids, weighted by class sizes.
  Indeed, the numerator of the criterion (Rayleigh ratio) is equal to the variance (inertia) of the projections $\mathbf{a}^\top \mathbf{g}_\ell$ with weights $\frac{n_\ell}{n}$:

$$\mathbf{a}^\top \mathbf{B}\mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{a}^\top \mathbf{g}_\ell \mathbf{g}_\ell^\top \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} (\mathbf{a}^\top \mathbf{g}_\ell)^2.$$

  The denominator is $\mathbf{a}^\top \mathbf{a}$ is the squared norm of $\mathbf{a}$. Hence, $\mathbf{a}_1$ is found by solving the PCA problem:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{I}_p \mathbf{a}} = \max_{\mathbf{a}, \|\mathbf{a}\|=1} I_a(\mathbf{g}_1, \dots, \mathbf{g}_m).$$

Main result (proof)

- First observe that when $\mathbf{W} = \mathbf{I}_p$, then LDA = PCA on the centroids, weighted by class sizes.
  Indeed, the numerator of the criterion (Rayleigh ratio) is equal to the variance (inertia) of the projections $\mathbf{a}^\top \mathbf{g}_\ell$ with weights $\frac{n_\ell}{n}$:

$$\mathbf{a}^\top \mathbf{B} \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} \mathbf{a}^\top \mathbf{g}_\ell \mathbf{g}_\ell^\top \mathbf{a} = \sum_{\ell=1}^{m} \frac{n_\ell}{n} (\mathbf{a}^\top \mathbf{g}_\ell)^2.$$

  The denominator is $\mathbf{a}^\top \mathbf{a}$ is the squared norm of $\mathbf{a}$. Hence, $\mathbf{a}_1$ is found by solving the PCA problem:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^\top \mathbf{B} \mathbf{a}}{\mathbf{a}^\top \mathbf{I}_p \mathbf{a}} = \max_{\mathbf{a}, \|\mathbf{a}\|=1} I_a(\mathbf{g}_1, \dots, \mathbf{g}_m).$$

  The same is true for $\mathbf{a}_2, \dots, \mathbf{a}_m$, since $\mathbf{W}$-orthog. = orthog.

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \quad \to$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \quad \to \quad \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{x}_i \quad \to \quad \frac{1}{n_\ell} \sum_{i \in \Omega_\ell} \mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

$$\mathbf{x}_i - \mathbf{g}_\ell \quad \to$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{x}_i \quad \to \quad \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

$$\mathbf{x}_i - \mathbf{g}_\ell \quad \to \quad \mathbf{W}^{-1/2}(\mathbf{x}_i - \mathbf{g}_\ell)$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{x}_i \quad \to \quad \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

$$\mathbf{x}_i - \mathbf{g}_\ell \quad \to \quad \mathbf{W}^{-1/2}(\mathbf{x}_i - \mathbf{g}_\ell)$$

$$\mathbf{W} = \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top \quad \to$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{x}_i \quad \to \quad \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

$$\mathbf{x}_i - \mathbf{g}_\ell \quad \to \quad \mathbf{W}^{-1/2}(\mathbf{x}_i - \mathbf{g}_\ell)$$

$$\mathbf{W} = \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top \quad \to \quad \mathbf{W}^{-1/2}(\mathbf{W})\mathbf{W}^{-1/2} = \mathbf{I}_\mathbf{p}$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{x}_i \quad \to \quad \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

$$\mathbf{x}_i - \mathbf{g}_\ell \quad \to \quad \mathbf{W}^{-1/2}(\mathbf{x}_i - \mathbf{g}_\ell)$$

$$\mathbf{W} = \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top \quad \to \quad \mathbf{W}^{-1/2}(\mathbf{W})\mathbf{W}^{-1/2} = \mathbf{I_p}$$

$$\mathbf{B} = \sum_{\ell=1}^{m}\frac{n_\ell}{n}\mathbf{g}_\ell\mathbf{g}_\ell^\top \quad \to$$

Main result (proof)

- Then, the idea is to sphere the data in order to have a identity within-class covariance matrix. This is obtained with $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$:

$$\mathbf{g}_\ell = \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{x}_i \quad\to\quad \frac{1}{n_\ell}\sum_{i\in\Omega_\ell}\mathbf{W}^{-1/2}\mathbf{x}_i = \mathbf{W}^{-1/2}\mathbf{g}_\ell$$

$$\mathbf{x}_i - \mathbf{g}_\ell \quad\to\quad \mathbf{W}^{-1/2}(\mathbf{x}_i - \mathbf{g}_\ell)$$

$$\mathbf{W} = \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)^\top \quad\to\quad \mathbf{W}^{-1/2}(\mathbf{W})\mathbf{W}^{-1/2} = \mathbf{I}_\mathbf{p}$$

$$\mathbf{B} = \sum_{\ell=1}^{m}\frac{n_\ell}{n}\mathbf{g}_\ell\mathbf{g}_\ell^\top \quad\to\quad \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$$

Main result (proof)

- Consequently, PCA for the sphered centroids is written

$$\mathbf{a}_1^* \quad = \quad \text{argmax}_{\mathbf{a}^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*}$$

Main result (proof)

- Consequently, PCA for the sphered centroids is written

$$\begin{aligned} \mathbf{a}_1^* &= \mathrm{argmax}_{\mathbf{a}^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*} \\ \mathbf{a}_2^* &= \mathrm{argmax}_{\mathbf{a}^*, \mathbf{a}^* \perp \mathbf{a}_1^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*} \qquad \ldots \end{aligned}$$

Main result (proof)

- Consequently, PCA for the sphered centroids is written

$$\mathbf{a}_1^* = \mathrm{argmax}_{\mathbf{a}^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*}$$

$$\mathbf{a}_2^* = \mathrm{argmax}_{\mathbf{a}^*, \mathbf{a}^* \perp \mathbf{a}_1^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*} \qquad \dots$$

- Now, reparametrize this optimization problem on $\mathbf{a}^*$ with $\mathbf{a} = \mathbf{W}^{-1/2}\mathbf{a}^*$. This gives the LDA problem:

Main result (proof)

- Consequently, PCA for the sphered centroids is written

$$\mathbf{a}_1^* = \text{argmax}_{\mathbf{a}^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*}$$

$$\mathbf{a}_2^* = \text{argmax}_{\mathbf{a}^*, \mathbf{a}^* \perp \mathbf{a}_1^*} \frac{\mathbf{a}^{*\top}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{\mathbf{a}^{*\top}\mathbf{a}^*} \qquad \ldots$$

- Now, reparametrize this optimization problem on $\mathbf{a}^*$ with $\mathbf{a} = \mathbf{W}^{-1/2}\mathbf{a}^*$. This gives the LDA problem:

$$\mathbf{a}_1 = \text{argmax}_{\mathbf{a}} \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}}$$

Main result (proof)

- Consequently, PCA for the sphered centroids is written

$$\begin{aligned}
\mathbf{a}_1^* &= \mathrm{argmax}_{\mathbf{a}^*} \frac{{\mathbf{a}^*}^\top (\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{{\mathbf{a}^*}^\top \mathbf{a}^*} \\
\mathbf{a}_2^* &= \mathrm{argmax}_{\mathbf{a}^*, \mathbf{a}^* \perp \mathbf{a}_1^*} \frac{{\mathbf{a}^*}^\top (\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{a}^*}{{\mathbf{a}^*}^\top \mathbf{a}^*} \qquad \ldots
\end{aligned}$$

- Now, reparametrize this optimization problem on $\mathbf{a}^*$ with $\mathbf{a} = \mathbf{W}^{-1/2}\mathbf{a}^*$. This gives the LDA problem:

$$\begin{aligned}
\mathbf{a}_1 &= \mathrm{argmax}_{\mathbf{a}} \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}} \\
\mathbf{a}_2 &= \mathrm{argmax}_{\mathbf{a}, \mathbf{a} \perp_{\mathbf{W}} \mathbf{a}_1} \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}} \ldots
\end{aligned}$$

Remarks

- In the textbook $\mathbf{B}, \mathbf{W}$ are denoted $\mathbf{S}_e, \mathbf{S}_r$, in order to emphasize that they correspond to *estimators* (of unknown proba. objects).

Remarks

- In the textbook $\mathbf{B}, \mathbf{W}$ are denoted $\mathbf{S}_e, \mathbf{S}_r$, in order to emphasize that they correspond to *estimators* (of unknown proba. objects).
- Link between the diagonalization of the symmetric matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ and the diagonalization of the matrix $\mathbf{B}\mathbf{W}^{-1}$:

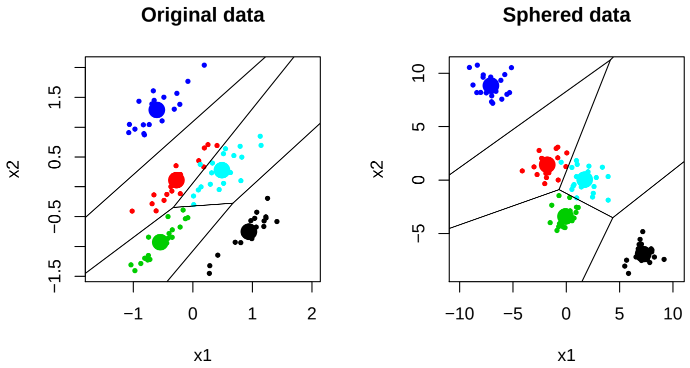
$$\begin{aligned} \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{a}^* = \lambda\mathbf{a}^* \quad &\Leftrightarrow \quad \mathbf{B}\mathbf{W}^{-1/2}\mathbf{a}^* = \lambda\mathbf{W}^{1/2}\mathbf{a}^* \\ &\Leftrightarrow \quad \mathbf{B}\mathbf{W}^{-1}(\mathbf{W}^{1/2}\mathbf{a}^*) = \lambda(\mathbf{W}^{1/2}\mathbf{a}^*) \end{aligned}$$

LDA, exploration: Recap

- LDA finds linear combinations of coordinates that maximize the between-class variance relatively to the within-class variance.
- LDA is equivalent to do PCA of the centroids with Mahalanobis metric, i.e. PCA on sphered centroids.

LDA as a classification tool: Theory

Case of classes of equal sizes



Visualization of the linear frontiers for LDA in the 2D case, when classes have the same size. For sphered data, it is the Voronoi tesselation.

Case of classes of equal sizes

When all classes have the same size, the optimal rule (see next slides) for classification is to predict by the closest centroid for sphered data:

$$\text{For a given } \mathbf{x}, \text{ choose } \ell \text{ such that } \delta(\ell) = \|\mathbf{W}^{-1/2}\mathbf{x} - \mathbf{W}^{-1/2}\mathbf{g}_\ell\| \text{ is minimal.}$$

Case of classes of equal sizes

When all classes have the same size, the optimal rule (see next slides) for classification is to predict by the closest centroid for sphered data:

$$\text{For a given } \mathbf{x}, \text{ choose } \ell \text{ such that } \delta(\ell) = \|\mathbf{W}^{-1/2}\mathbf{x} - \mathbf{W}^{-1/2}\mathbf{g}_\ell\| \text{ is minimal.}$$

This gives the Voronoi tesselation of centroids in the sphered space:

$$\text{For a given } \mathbf{x}, \text{ and given } \ell_1, \ell_2, \text{ prefer } \ell_1 \text{ to } \ell_2 \text{ if } \delta(\ell_1) \leq \delta(\ell_2).$$

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).
- As $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$ is linear, frontiers remain linear in the original space. This justifies the name "Linear" discriminant analysis.

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).
- As $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$ is linear, frontiers remain linear in the original space. This justifies the name "Linear" discriminant analysis.
- Let us compute the hyperplane equations in the original space.

$$\delta(\ell_1)^2 \quad = \quad \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_1})\|^2 =$$

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).
- As $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$ is linear, frontiers remain linear in the original space. This justifies the name "Linear" discriminant analysis.
- Let us compute the hyperplane equations in the original space.

$$\delta(\ell_1)^2 = \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_1})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_1}$$

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).
- As $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$ is linear, frontiers remain linear in the original space. This justifies the name "Linear" discriminant analysis.
- Let us compute the hyperplane equations in the original space.

$$\begin{aligned}
\delta(\ell_1)^2 &= \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_1})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_1} \\
\delta(\ell_2)^2 &= \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_2})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_2}
\end{aligned}$$

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).
- As $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$ is linear, frontiers remain linear in the original space. This justifies the name "Linear" discriminant analysis.
- Let us compute the hyperplane equations in the original space.

$$\begin{aligned}
\delta(\ell_1)^2 &= \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_1})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_1} \\
\delta(\ell_2)^2 &= \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_2})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_2}
\end{aligned}$$

When writing the equation $\delta(\ell_1) = \delta(\ell_2)$, the quadratic term disappear, and:

$$\delta(\ell_1)^2 = \delta(\ell_2)^2 \quad \Leftrightarrow$$

Case of classes of equal sizes

- In the sphered space, prediction frontiers are linear (defined by bisector hyperplanes).
- As $\mathbf{x} \to \mathbf{W}^{-1/2}\mathbf{x}$ is linear, frontiers remain linear in the original space. This justifies the name "Linear" discriminant analysis.
- Let us compute the hyperplane equations in the original space.

$$\begin{aligned}
\delta(\ell_1)^2 &= \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_1})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_1} \\
\delta(\ell_2)^2 &= \|\mathbf{W}^{-1/2}(\mathbf{x} - \mathbf{g}_{\ell_2})\|^2 = \mathbf{x}^\top \mathbf{W}^{-1}\mathbf{x} - 2\mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{x} + \mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_2}
\end{aligned}$$

When writing the equation $\delta(\ell_1) = \delta(\ell_2)$, the quadratic term disappear, and:

$$\delta(\ell_1)^2 = \delta(\ell_2)^2 \quad \Leftrightarrow \quad 2(\mathbf{g}_{\ell_1} - \mathbf{g}_{\ell_2})^\top \mathbf{W}^{-1}\mathbf{x} = \mathbf{g}_{\ell_1}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_1} - \mathbf{g}_{\ell_2}^\top \mathbf{W}^{-1}\mathbf{g}_{\ell_2}$$

General case, probabilistic approach

In the general case, we need to rely on a more probabilistic approach. We consider the following Gaussian mixture model.

Let G a discrete random variables on $\{1, \dots, m\}$ with $P(G = \ell) = \pi_\ell$. Let $\mathbf{X}$ a random vector of $\mathbb{R}^p$, such that

$$\mathbf{X}|G = \ell \sim \mathcal{N}(\mathbf{g}_\ell, \mathbf{W}_\ell)$$

with $g_\ell \in \mathbb{R}^p$ and $\mathbf{W}_\ell$ a covariance matrix ($\ell = 1, \dots, m$).

General case, probabilistic approach

In the general case, we need to rely on a more probabilistic approach. We consider the following Gaussian mixture model.

Let G a discrete random variables on $\{1, \ldots, m\}$ with $P(G = \ell) = \pi_\ell$. Let $\mathbf{X}$ a random vector of $\mathbb{R}^p$, such that

$$\mathbf{X}|G = \ell \sim \mathcal{N}(\mathbf{g}_\ell, \mathbf{W}_\ell)$$

with $g_\ell \in \mathbb{R}^p$ and $\mathbf{W}_\ell$ a covariance matrix ($\ell = 1, \ldots, m$).

<u>Exercice.</u> Show that $\mathbf{X}$ has the density $f_{\mathbf{X}}(\mathbf{x}) = \sum_{\ell=1}^m \pi_\ell f_{\mathbf{X}|G=\ell}(\mathbf{x})$.

General case, probabilistic approach

- Recall the Bayes classifier optimal rule for probabilistic models:

$$\text{For a given } \mathbf{x}, \text{ choose } \ell \text{ such that } P(G = \ell | \mathbf{X} = x) \text{ is maximal.}$$

General case, probabilistic approach

- Recall the Bayes classifier optimal rule for probabilistic models:

$$\text{For a given } \mathbf{x},\ \text{choose } \ell \text{ such that } P(G = \ell | \mathbf{X} = x) \text{ is maximal.}$$

- Reminder: Bayes theorem, when $P(A) \neq 0$: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$.

General case, probabilistic approach

- Recall the Bayes classifier optimal rule for probabilistic models:

$$\text{For a given } \mathbf{x}, \text{ choose } \ell \text{ such that } P(G = \ell | \mathbf{X} = x) \text{ is maximal.}$$

- Reminder: Bayes theorem, when $P(A) \neq 0$: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$.

- In our context (G discrete, X continuous), Bayes theorem is:

$$P(G = \ell | \mathbf{X} = x) = \frac{f_{\mathbf{X}|G=\ell}(\mathbf{x}) P(G = \ell)}{f_{\mathbf{X}}(\mathbf{x})}$$

General case, probabilistic approach

- Thus the classification rule for the Gaussian mixture model is:

$$\text{For a given } \mathbf{x}, \text{ choose } \ell \text{ such that } f_{\mathbf{X}|G=\ell}(\mathbf{x})\pi_\ell \text{ is maximal.}$$

General case, probabilistic approach

- Thus the classification rule for the Gaussian mixture model is:

$$\text{For a given } \mathbf{x}, \text{ choose } \ell \text{ such that } f_{\mathbf{X}|G=\ell}(\mathbf{x})\pi_\ell \text{ is maximal.}$$

- Equivalently, this defines a tesselation of the space:

$$\begin{aligned}&\text{For a given } \mathbf{x}, \text{ and given } \ell_1, \ell_2,\\ &\text{Prefer } \ell_1 \text{ to } \ell_2 \quad \text{if} \quad f_{\mathbf{X}|G=\ell_1}(\mathbf{x})\pi_{\ell_1} \geq f_{\mathbf{X}|G=\ell_2}(\mathbf{x})\pi_{\ell_2}.\end{aligned}$$

General case, probabilistic approach

- Now, we have

$$f_{\mathbf{X}|G=\ell}(\mathbf{x}) \quad = \quad (2\pi)^{-d/2}|\mathbf{W}_\ell|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell)}$$

General case, probabilistic approach

- Now, we have

$$\begin{aligned} f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= (2\pi)^{-d/2}|\mathbf{W}_\ell|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell)} \\ -2\log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d\log(2\pi) + \log|\mathbf{W}_\ell| + (\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell) \end{aligned}$$

  which is, as a function of **x**, a...

General case, probabilistic approach

- Now, we have

$$\begin{aligned} f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= (2\pi)^{-d/2}|\mathbf{W}_\ell|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell)} \\ -2\log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d\log(2\pi) + \log|\mathbf{W}_\ell| + (\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell) \end{aligned}$$

which is, as a function of $\mathbf{x}$, a... quadratic polynomial.

General case, probabilistic approach

- Now, we have

$$\begin{aligned} f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= (2\pi)^{-d/2}|\mathbf{W}_\ell|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{g}_\ell)^\top\mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell)} \\ -2\log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d\log(2\pi)+\log|\mathbf{W}_\ell|+(\mathbf{x}-\mathbf{g}_\ell)^\top\mathbf{W}_\ell^{-1}(\mathbf{x}-\mathbf{g}_\ell) \end{aligned}$$

  which is, as a function of $\mathbf{x}$, a... quadratic polynomial.

- Thus, in general, the classification rule gives quadratic frontiers. This is quadratic discriminant analysis.

$$\begin{aligned} f_{\mathbf{X}|G=\ell_1}(\mathbf{x})\pi_{\ell_1} &= f_{\mathbf{X}|G=\ell_2}(\mathbf{x})\pi_{\ell_2} \\ -\log(f_{\mathbf{X}|G=\ell_1}(\mathbf{x}))-\log(\pi_{\ell_1}) &= -\log(f_{\mathbf{X}|G=\ell_2}(\mathbf{x}))-\log(\pi_{\ell_2}). \end{aligned}$$

General case, probabilistic approach

- Now, assume homoscedasticity: $\mathbf{W}_\ell = \mathbf{W}$ for all ℓ. Then:

General case, probabilistic approach

- Now, assume homoscedasticity: $\mathbf{W}_\ell = \mathbf{W}$ for all ℓ. Then:

$$-2 \log f_{\mathbf{X}|G=\ell}(\mathbf{x}) \quad = \quad d \log(2\pi) + \log|\mathbf{W}| + (\mathbf{x} - \mathbf{g}_\ell)^\top \mathbf{W}^{-1}(\mathbf{x} - \mathbf{g}_\ell)$$

General case, probabilistic approach

- Now, assume homoscedasticity: $\mathbf{W}_\ell = \mathbf{W}$ for all ℓ. Then:

$$\begin{aligned} -2 \log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d \log(2\pi) + \log |\mathbf{W}| + (\mathbf{x} - \mathbf{g}_\ell)^\top \mathbf{W}^{-1} (\mathbf{x} - \mathbf{g}_\ell) \\ &= d \log(2\pi) + \log |\mathbf{W}| + \delta(\ell)^2 \end{aligned}$$

General case, probabilistic approach

- Now, assume homoscedasticity: $\mathbf{W}_\ell = \mathbf{W}$ for all ℓ. Then:

$$
\begin{aligned}
-2 \log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d \log(2\pi) + \log|\mathbf{W}| + (\mathbf{x} - \mathbf{g}_\ell)^\top \mathbf{W}^{-1} (\mathbf{x} - \mathbf{g}_\ell) \\
&= d \log(2\pi) + \log|\mathbf{W}| + \delta(\ell)^2
\end{aligned}
$$

- Now, Bayes classif. rule simplifies to linear discriminant analysis:

$$
f_{\mathbf{X}|G=\ell_1}(\mathbf{x})\pi_{\ell_1} = f_{\mathbf{X}|G=\ell_2}(\mathbf{x})\pi_{\ell_2}
$$

General case, probabilistic approach

- Now, assume homoscedasticity: $\mathbf{W}_\ell = \mathbf{W}$ for all ℓ. Then:

$$\begin{aligned} -2\log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d\log(2\pi) + \log|\mathbf{W}| + (\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}^{-1}(\mathbf{x}-\mathbf{g}_\ell) \\ &= d\log(2\pi) + \log|\mathbf{W}| + \delta(\ell)^2 \end{aligned}$$

- Now, Bayes classif. rule simplifies to linear discriminant analysis:

$$\begin{aligned} f_{\mathbf{X}|G=\ell_1}(\mathbf{x})\pi_{\ell_1} &= f_{\mathbf{X}|G=\ell_2}(\mathbf{x})\pi_{\ell_2} \\ -2\log(f_{\mathbf{X}|G=\ell_1}(\mathbf{x})) - 2\log(\pi_{\ell_1}) &= -2\log(f_{\mathbf{X}|G=\ell_2}(\mathbf{x})) - 2\log(\pi_{\ell_2}) \end{aligned}$$

General case, probabilistic approach

- Now, assume homoscedasticity: $\mathbf{W}_\ell = \mathbf{W}$ for all ℓ. Then:

$$\begin{aligned} -2\log f_{\mathbf{X}|G=\ell}(\mathbf{x}) &= d\log(2\pi) + \log|\mathbf{W}| + (\mathbf{x}-\mathbf{g}_\ell)^\top \mathbf{W}^{-1}(\mathbf{x}-\mathbf{g}_\ell) \\ &= d\log(2\pi) + \log|\mathbf{W}| + \delta(\ell)^2 \end{aligned}$$

- Now, Bayes classif. rule simplifies to linear discriminant analysis:

$$\begin{aligned} f_{\mathbf{X}|G=\ell_1}(\mathbf{x})\pi_{\ell_1} &= f_{\mathbf{X}|G=\ell_2}(\mathbf{x})\pi_{\ell_2} \\ -2\log(f_{\mathbf{X}|G=\ell_1}(\mathbf{x})) - 2\log(\pi_{\ell_1}) &= -2\log(f_{\mathbf{X}|G=\ell_2}(\mathbf{x})) - 2\log(\pi_{\ell_2}) \\ \delta(\ell_1)^2 - 2\log(\pi_{\ell_1}) &= \delta(\ell_2)^2 - 2\log(\pi_{\ell_2}) \end{aligned}$$

LDA, prediction: Recap

- In general, Bayes rule gives quadratic prediction frontiers
  $\rightarrow$ quadratic discriminant analysis
- Under homoscedasticity, frontiers become linear
  $\rightarrow$ linear discriminant analysis
- For LDA, the rule is to choose the closest centroid for sphered data, enhanced by the term $-2\log(\pi_\ell)$, linked to class size. When π_ℓ does not depend on ℓ, it comes down to choose the closest centroid for shered data.