



copyright (c) Nalisa Prints 2001

A Geometry of Data Sets

- Adi Ben-Israel (Rutgers University, USA)
- Yuri Levin (Queen's University, Canada)

A Geometry of Data Sets

- Adi Ben-Israel (Rutgers University, USA)
- Yuri Levin (Queen's University, Canada)
- Cem Iyigun (Rutgers Univ.)
- Zachary Stoumbos (Rutgers Univ.)

A Geometry of Data Sets

- Adi Ben-Israel (Rutgers University, USA)
- Yuri Levin (Queen's University, Canada)
- Cem Iyigun (Rutgers Univ.)
- Zachary Stoumbos (Rutgers Univ.)

Matrix Theory Conference, Haifa, January 2005

We thank the organizers

Statistical Learning

The objects of study are vectors $\mathbf{v} = \begin{bmatrix} \mathbf{x} \\ y \end{bmatrix} \in \mathbb{R}^{p+1}$

$\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^p$ (**inputs, attributes**) observable, readily measurable.

$y \in Y \subset \mathbb{R}$ (**output, class**) more difficult to measure.

Statistical Learning

The objects of study are vectors $\mathbf{v} = \begin{bmatrix} \mathbf{x} \\ y \end{bmatrix} \in \mathbb{R}^{p+1}$

$\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^p$ (**inputs, attributes**) observable, readily measurable.

$y \in Y \subset \mathbb{R}$ (**output, class**) more difficult to measure.

Problem: Predict (or estimate) y given \mathbf{x} .

Data: N given observations (**data set**)

$$\mathbf{D} = \{(\mathbf{x}_i, y_i) : i = 1, \dots, N\}$$

Statistical Learning

The objects of study are vectors $\mathbf{v} = \begin{bmatrix} \mathbf{x} \\ y \end{bmatrix} \in \mathbb{R}^{p+1}$

$\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^p$ (**inputs, attributes**) observable, readily measurable.

$y \in Y \subset \mathbb{R}$ (**output, class**) more difficult to measure.

Problem: Predict (or estimate) y given \mathbf{x} .

Data: N given observations (**data set**)

$$\mathbf{D} = \{(\mathbf{x}_i, y_i) : i = 1, \dots, N\}$$

Procedure:

1. Select subset $\mathbf{T} \subset \mathbf{D}$ (**training set**)
2. Use \mathbf{T} to determine a **rule** $f : \mathbf{x} \rightarrow y$

$$y = f(\mathbf{x})$$

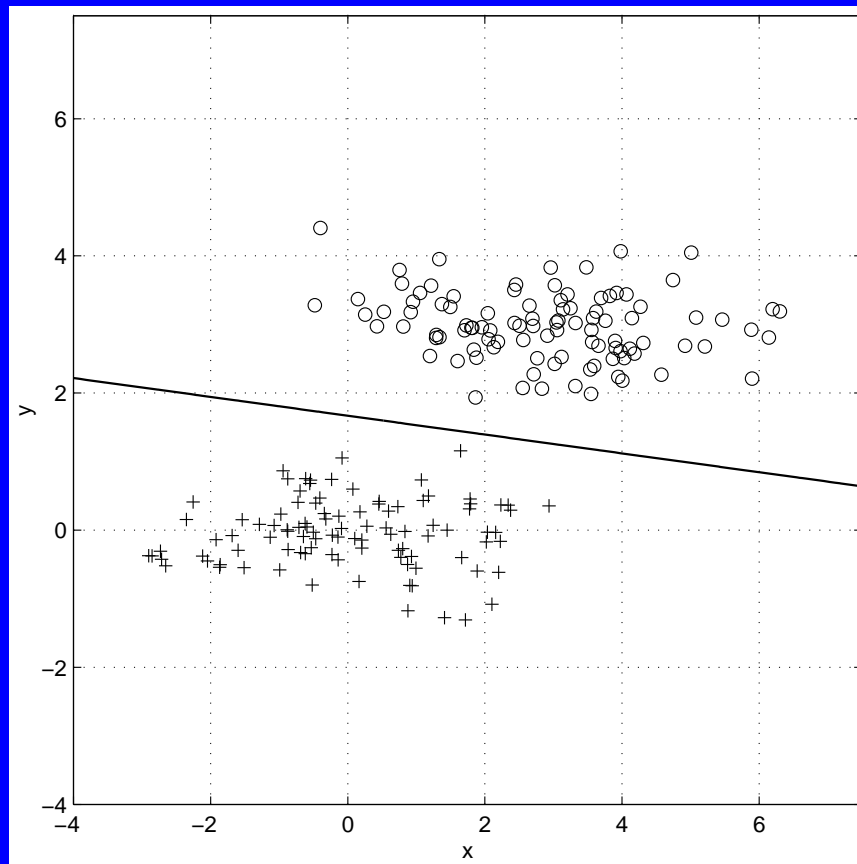
3. Test the performance of f on $\mathbf{D} \setminus \mathbf{T}$

A linear discriminant rule

- $x \sim N(3, 1.5)$ + $x \sim N(0, 1.5)$
- $y \sim N(3, 0.5)$ + $y \sim N(0, 0.5)$

A linear discriminant rule

- $x \sim N(3, 1.5)$ + $x \sim N(0, 1.5)$
- $y \sim N(3, 0.5)$ + $y \sim N(0, 0.5)$

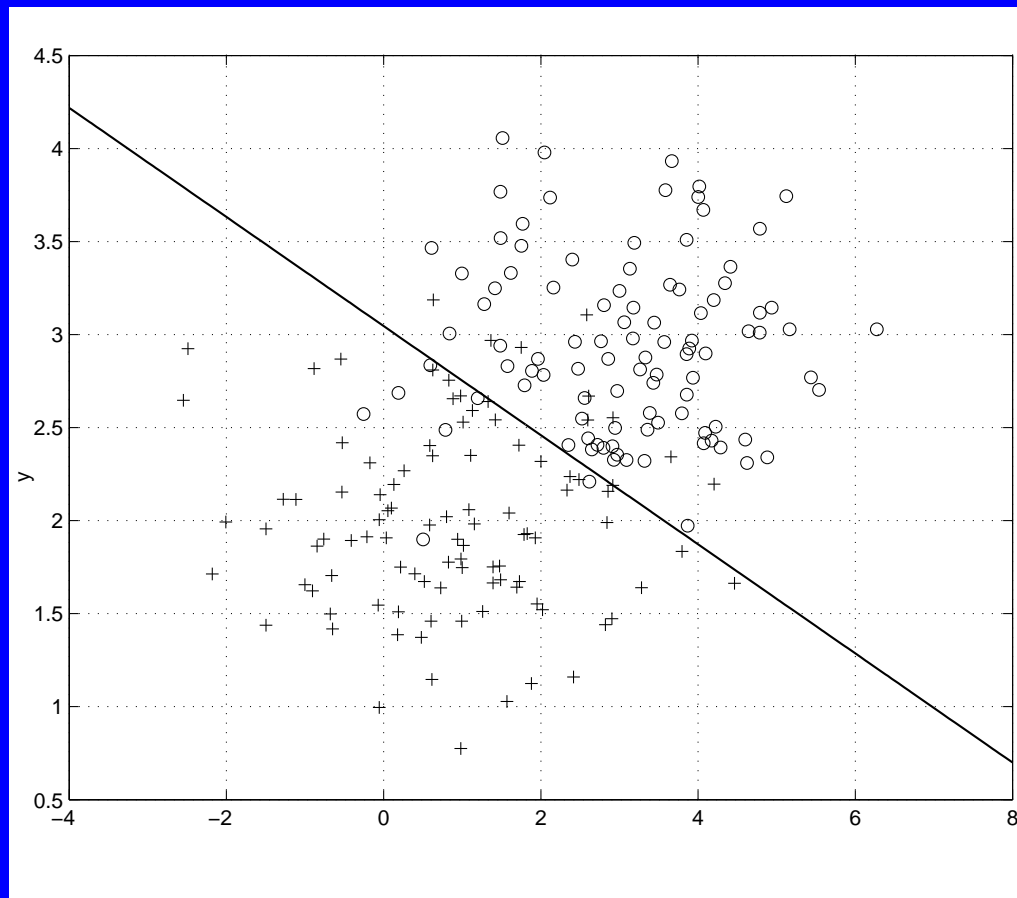


A linear discriminant rule

- $x \sim N(3, 1.5)$ + $x \sim N(1, 1.5)$
- $y \sim N(3, 0.5)$ + $y \sim N(2, 0.5)$

A linear discriminant rule

- $x \sim N(3, 1.5)$
- $y \sim N(3, 0.5)$
- + $x \sim N(1, 1.5)$
- + $y \sim N(2, 0.5)$



Medical applications

Typically: $\mathbf{x} = (x_1, \dots, x_p)$ results of **diagnostic tests**,
 $y \in \{0, 1\}$ denoting respectively the absence or presence of **disease**.

Medical applications

Typically: $\mathbf{x} = (x_1, \dots, x_p)$ results of **diagnostic tests**,
 $y \in \{0, 1\}$ denoting respectively the absence or presence of **disease**.

y dictates the course of treatment.

The two possible errors:

- type 1: **false positive**, and
- type 2: **false negative**, differ in their consequences.

Medical applications

Typically: $\mathbf{x} = (x_1, \dots, x_p)$ results of **diagnostic tests**,
 $y \in \{0, 1\}$ denoting respectively the absence or presence of **disease**.

y dictates the course of treatment.

The two possible errors:

- type 1: **false positive**, and
- type 2: **false negative**, differ in their consequences.

C. Merz and P. Murphy, *UCI Repository of machine learning databases*.

Dept. of Info. and Comp. Sci., Univ. of California, Irvine, CA, 1996.

www.ics.uci.edu/~mlearn/MLRepository.html

Medical applications

Typically: $\mathbf{x} = (x_1, \dots, x_p)$ results of **diagnostic tests**,
 $y \in \{0, 1\}$ denoting respectively the absence or presence of **disease**.

y dictates the course of treatment.

The two possible errors:

- type 1: **false positive**, and
- type 2: **false negative**, differ in their consequences.

C. Merz and P. Murphy, *UCI Repository of machine learning databases*.

Dept. of Info. and Comp. Sci., Univ. of California, Irvine, CA, 1996.

www.ics.uci.edu/~mlearn/MLRepository.html

T. Lim, W. Loh, and Y. Shih, *A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms*, *Machine Learning* **40**(2000), 203–228

A Naive Proposal

1. Define a distance function $d : \mathbf{X} \times Y \rightarrow \mathbb{R}$, e.g.,

$$d\left(\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}$$

A Naive Proposal

1. Define a distance function $d : \mathbf{X} \times Y \rightarrow \mathbb{R}$, e.g.,

$$d\left(\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}$$

2. Use d for classification of \mathcal{D} in **clusters** $\{\Omega_1, \dots, \Omega_m\}$.

A Naive Proposal

1. Define a distance function $d : \mathbf{X} \times Y \rightarrow \mathbb{R}$, e.g.,

$$d\left(\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}$$

2. Use d for classification of \mathcal{D} in **clusters** $\{\Omega_1, \dots, \Omega_m\}$.
3. For each cluster Ω_i compute:
 - a **center** \bar{y}_i of y ,
 - Ω_i^X the X -projection of Ω_i

A Naive Proposal

1. Define a distance function $d : \mathbf{X} \times Y \rightarrow \mathbb{R}$, e.g.,

$$d\left(\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}$$

2. Use d for classification of \mathcal{D} in **clusters** $\{\Omega_1, \dots, \Omega_m\}$.
3. For each cluster Ω_i compute:
 - a **center** \bar{y}_i of y ,
 - Ω_i^X the X -projection of Ω_i
4. Given $\mathbf{x} \in X$, determine the nearest projected cluster, say Ω_i^X .

A Naive Proposal

1. Define a distance function $d : \mathbf{X} \times Y \rightarrow \mathbb{R}$, e.g.,

$$d\left(\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}$$

2. Use d for classification of \mathcal{D} in **clusters** $\{\Omega_1, \dots, \Omega_m\}$.
3. For each cluster Ω_i compute:
 - a **center** \bar{y}_i of y ,
 - Ω_i^X the X -projection of Ω_i
4. Given $\mathbf{x} \in X$, determine the nearest projected cluster, say Ω_i^X .
5. Use \bar{y}_i as estimate for y .

A Naive Proposal

1. Define a distance function $d : \mathbf{X} \times Y \rightarrow \mathbb{R}$, e.g.,

$$d\left(\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(\mathbf{x}_1, \mathbf{x}_2) + \alpha d_Y^2(y_1, y_2)}$$

2. Use d for classification of \mathcal{D} in **clusters** $\{\Omega_1, \dots, \Omega_m\}$.
3. For each cluster Ω_i compute:
 - a **center** \bar{y}_i of y ,
 - Ω_i^X the X -projection of Ω_i
4. Given $\mathbf{x} \in X$, determine the nearest projected cluster, say Ω_i^X .
5. Use \bar{y}_i as estimate for y .

Yuri Levin and A. B-I, *Opsearch*, 2000

Name of Data Set	% Correct Predictions			% Errors		Lim et al	
	Mean	Max	Min	Type 1	Type 2	Max	Min
<i>Breast Cancer</i>	96.5	100	93.1	2.5	1.0	97	91
<i>Liver</i>	63.2	79.3	49.7	19.7	17.1	72	57
<i>Diabetes</i>	74.7	79.9	65.7	10.2	15.1	78	69
<i>Voting</i>	92.0	98.78	82.3	3.8	4.2	96	94
<i>Wine</i>	93.7	100	82.35	2.6	3.7	100	NA
<i>Hepatitis</i>	86.03	96.42	71.43	8.12	5.85	83	NA

Fisher's Discriminant: Separation of Populations with equal Covariances

Observations $\mathbf{x} \in \mathbb{R}^p$ from two populations with equal covariance Σ .

Fisher's Discriminant: Separation of Populations with equal Covariances

Observations $\mathbf{x} \in \mathbb{R}^p$ from two populations with equal covariance Σ .
Sample means $\bar{\mathbf{x}}_i$ and (pooled) sample variance S are computed.

Fisher's Discriminant: Separation of Populations with equal Covariances

Observations $\mathbf{x} \in \mathbb{R}^p$ from two populations with equal covariance Σ . Sample means $\bar{\mathbf{x}}_i$ and (pooled) sample variance S are computed.

It is required to find $\mathbf{a} \in \mathbb{R}^p$ maximizing

$$\frac{(\mathbf{a}^T \bar{\mathbf{x}}_1 - \mathbf{a}^T \bar{\mathbf{x}}_2)^2}{\mathbf{a}^T S \mathbf{a}}$$

Fisher's Discriminant: Separation of Populations with equal Covariances

Observations $\mathbf{x} \in \mathbb{R}^p$ from two populations with equal covariance Σ . Sample means $\bar{\mathbf{x}}_i$ and (pooled) sample variance S are computed.

It is required to find $\mathbf{a} \in \mathbb{R}^p$ maximizing

$$\frac{(\mathbf{a}^T \bar{\mathbf{x}}_1 - \mathbf{a}^T \bar{\mathbf{x}}_2)^2}{\mathbf{a}^T S \mathbf{a}}$$

Rationale: Let $y = \mathbf{a}^T \mathbf{x}$. Then

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\mathbf{a}^T \bar{\mathbf{x}}_1 - \mathbf{a}^T \bar{\mathbf{x}}_2)^2}{\mathbf{a}^T S \mathbf{a}}.$$

Fisher's Discriminant: Separation of Populations with equal Covariances

Let $\mathbf{d} := \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. The problem:

$$\max \{(\mathbf{a}^T \mathbf{d})^2 : \mathbf{a}^T S \mathbf{a} = 1\} \quad (\text{P})$$

Fisher's Discriminant: Separation of Populations with equal Covariances

Let $\mathbf{d} := \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. The problem:

$$\max \{(\mathbf{a}^T \mathbf{d})^2 : \mathbf{a}^T S \mathbf{a} = 1\} \quad (\text{P})$$

has the optimal solution

$$\mathbf{a} = \frac{1}{\sqrt{\mathbf{d}^T S^{-1} \mathbf{d}}} S^{-1} \mathbf{d}$$

Fisher's Discriminant: Separation of Populations with equal Covariances

Let $\mathbf{d} := \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. The problem:

$$\max \{(\mathbf{a}^T \mathbf{d})^2 : \mathbf{a}^T S \mathbf{a} = 1\} \quad (\text{P})$$

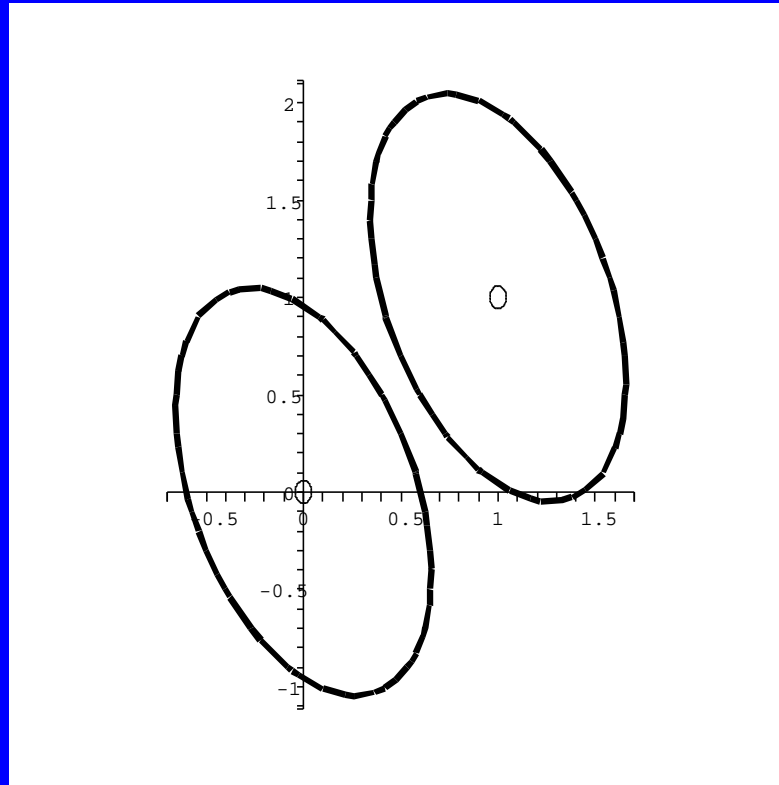
has the optimal solution

$$\mathbf{a} = \frac{1}{\sqrt{\mathbf{d}^T S^{-1} \mathbf{d}}} S^{-1} \mathbf{d}$$

and optimal value

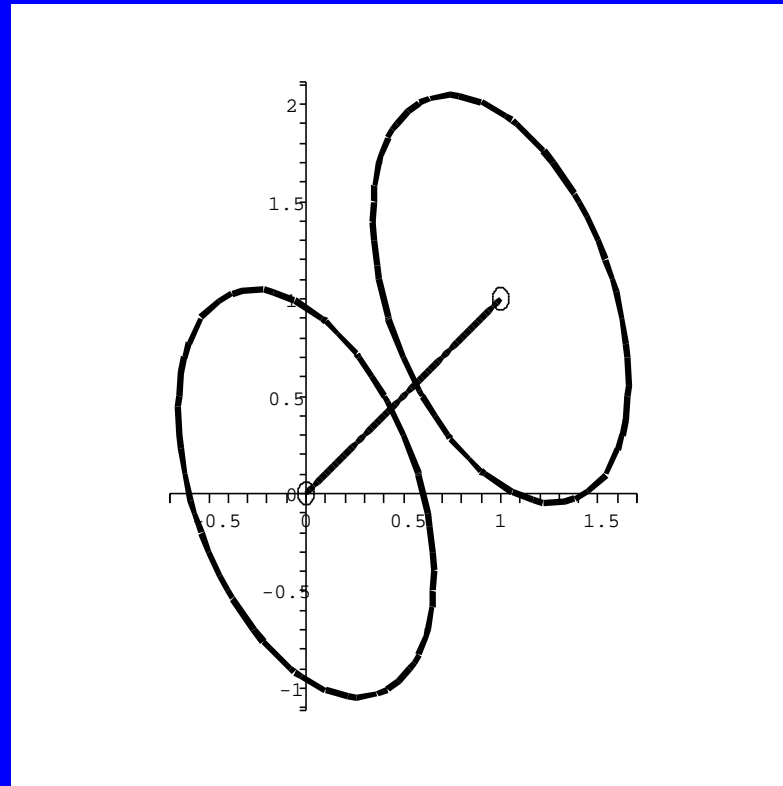
$$\max \frac{(\mathbf{a}^T \mathbf{d})^2}{\mathbf{a}^T S \mathbf{a}} = \mathbf{d}^T S^{-1} \mathbf{d}$$

Two populations $\sim \mathbf{N}(\mu_i, \Sigma)$, $i = 1, 2$

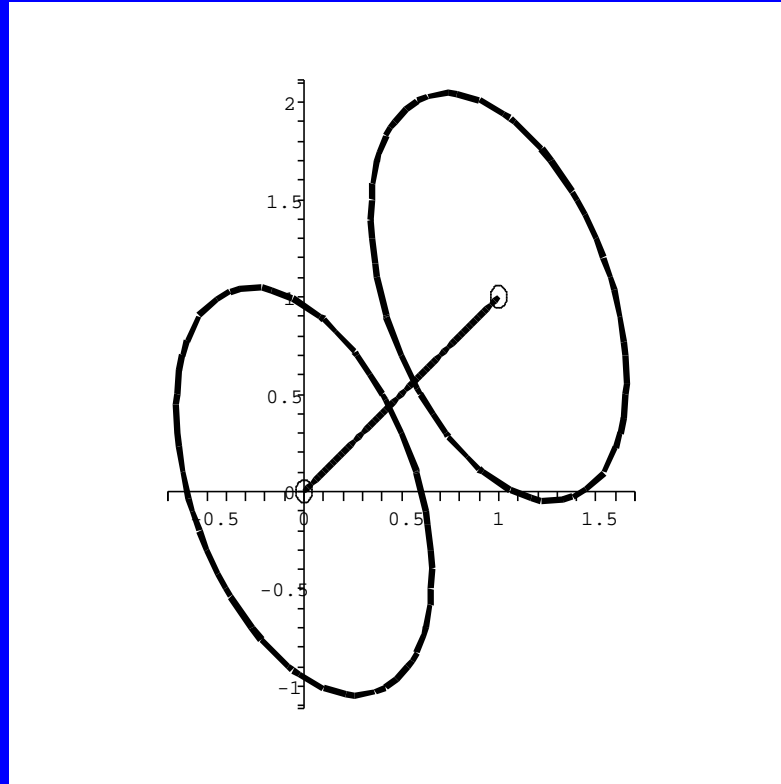


The samples represented by ellipses have means $\bar{\mathbf{x}}_i$, $i = 1, 2$ and variance S

$$\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$

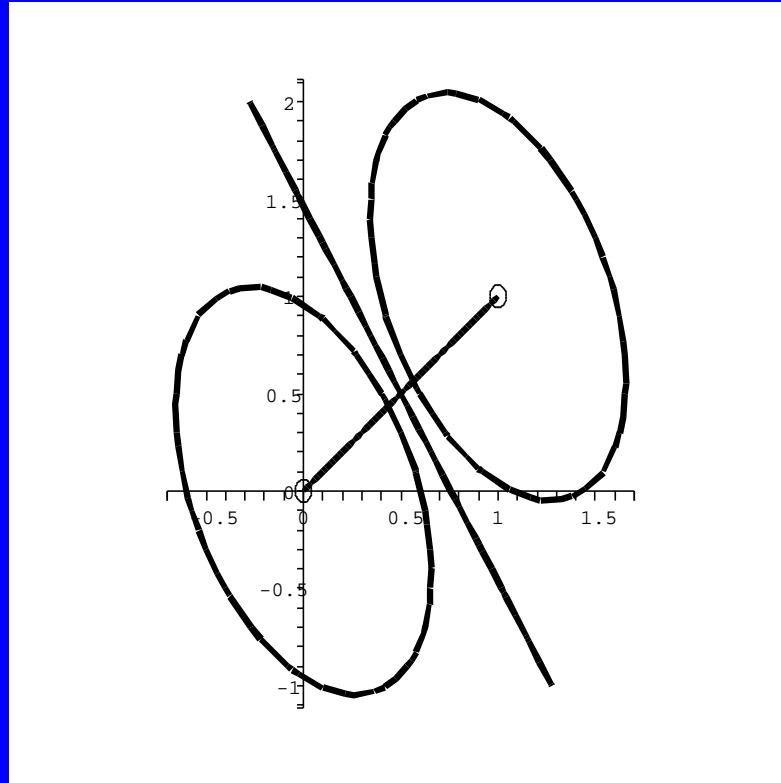


$$\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$



$$\max \frac{(\mathbf{a}^T \mathbf{d})^2}{\mathbf{a}^T S \mathbf{a}} \quad (\text{P})$$
$$\max \{ (\mathbf{a}^T \mathbf{d})^2 : \mathbf{a}^T S \mathbf{a} = 1 \}$$

The Fisher discriminant is given by the line $\mathbf{d}^T \mathbf{S}^{-1} \mathbf{x} = \alpha$



$$\alpha = \frac{1}{2} \mathbf{d}^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

Classification using Fisher's Discriminant

Classification using Fisher's Discriminant

Let $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, \mathbf{d} , S be as above.

Assign an observation \mathbf{x} to population 1 if

$$\mathbf{d}^T S^{-1} \mathbf{x} > \frac{1}{2} \mathbf{d}^T S^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

to population 2 otherwise.

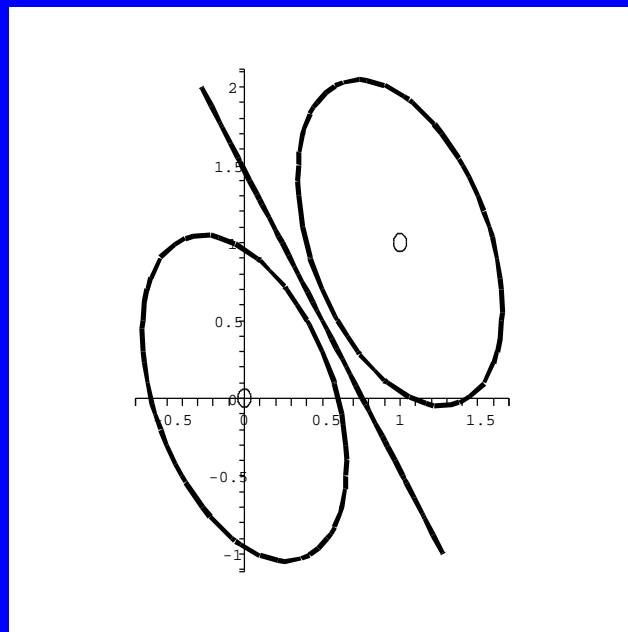
Classification using Fisher's Discriminant

Let $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, \mathbf{d} , S be as above.

Assign an observation \mathbf{x} to population 1 if

$$\mathbf{d}^T S^{-1} \mathbf{x} > \frac{1}{2} \mathbf{d}^T S^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

to population 2 otherwise.



An Optimization Problem

$\mathbf{d} \in \mathbb{R}^n$, $S \in \mathbb{R}^{n \times n}$ PSD.

The problem:

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\} \quad (\text{P})$$

An Optimization Problem

$\mathbf{d} \in \mathbb{R}^n$, $S \in \mathbb{R}^{n \times n}$ PSD.

The problem:

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\} \quad (\text{P})$$

Lagrangian:

$$L(\mathbf{x}, \lambda) = (\mathbf{d}^T \mathbf{x})^2 - \lambda(\mathbf{x}^T S \mathbf{x} - 1)$$

An Optimization Problem

$\mathbf{d} \in \mathbb{R}^n$, $S \in \mathbb{R}^{n \times n}$ PSD.

The problem:

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\} \quad (\text{P})$$

Lagrangian:

$$L(\mathbf{x}, \lambda) = (\mathbf{d}^T \mathbf{x})^2 - \lambda(\mathbf{x}^T S \mathbf{x} - 1)$$

An optimal solution must satisfy

$$\frac{1}{2} \nabla L(\mathbf{x}, \lambda) = (\mathbf{d}^T \mathbf{x})\mathbf{d} - \lambda S \mathbf{x} = \mathbf{0}$$

An Optimization Problem

$\mathbf{d} \in \mathbb{R}^n$, $S \in \mathbb{R}^{n \times n}$ PSD.

The problem:

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\} \quad (\text{P})$$

Lagrangian:

$$L(\mathbf{x}, \lambda) = (\mathbf{d}^T \mathbf{x})^2 - \lambda(\mathbf{x}^T S \mathbf{x} - 1)$$

An optimal solution must satisfy

$$\frac{1}{2} \nabla L(\mathbf{x}, \lambda) = (\mathbf{d}^T \mathbf{x})\mathbf{d} - \lambda S \mathbf{x} = \mathbf{0}$$

$$\therefore S \mathbf{x} = \left(\frac{\mathbf{d}^T \mathbf{x}}{\lambda} \right) \mathbf{d} \quad (1)$$

Case 1: $\mathbf{d} \in R(S)$

$$\mathbf{x} = \left(\frac{\mathbf{d}^T \mathbf{x}}{\lambda} \right) S^\dagger \mathbf{d} \quad (2)$$

$$\therefore \mathbf{x} = \alpha S^\dagger \mathbf{d}, \quad \alpha = \frac{\mathbf{d}^T \mathbf{x}}{\lambda} \quad (3)$$

$$\therefore \mathbf{x}^T S \mathbf{x} = \alpha^2 \mathbf{d}^T S^\dagger S S^\dagger \mathbf{d} = \alpha^2 \mathbf{d}^T S^\dagger \mathbf{d} = 1$$

$$\therefore \alpha^2 = \frac{1}{\mathbf{d}^T S^\dagger \mathbf{d}} \quad (4)$$

$$\therefore \mathbf{x} = \frac{1}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d}}} S^\dagger \mathbf{d} \quad (5)$$

An Optimization Problem (cont'd)

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\} \quad (\text{P})$$

The story so far: If $\mathbf{d} \in R(S)$ then

$$\mathbf{x} = \left(\frac{1}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d}}} \right) S^\dagger \mathbf{d} \quad (5)$$

$$(\mathbf{d}^T \mathbf{x})^2 = \left(\frac{\mathbf{d}^T S^\dagger \mathbf{d}}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d}}} \right)^2 = \mathbf{d}^T S^\dagger \mathbf{d} \quad (6)$$

Case 2: $\mathbf{d} \notin R(S)$ (so S is singular)

Let $\mathbf{z} = P_{N(S)} \mathbf{d} \quad \therefore \mathbf{z} \neq \mathbf{0}$

Let \mathbf{x}_0 satisfy

$$\mathbf{x}_0^T S \mathbf{x}_0 = 1$$

$$\mathbf{x}(t) := \mathbf{x}_0 + t\mathbf{z}$$

$$\therefore \mathbf{x}(t)^T S \mathbf{x}(t) = 1, \quad \forall t$$

An Optimization Problem (cont'd)

But

$$\begin{aligned}\mathbf{d}^T \mathbf{x}(t) &= \mathbf{d}^T \mathbf{x}_0 + t \mathbf{d}^T \mathbf{z} \\ &= \mathbf{d}^T \mathbf{x}_0 + t \mathbf{d}^T P_{N(S)} \mathbf{d} \\ &= \mathbf{d}^T \mathbf{x}_0 + t \|P_{N(S)} \mathbf{d}\|^2 \\ &= \mathbf{d}^T \mathbf{x}_0 + t \|\mathbf{z}\|^2\end{aligned}$$

$$\therefore |\mathbf{d}^T \mathbf{x}(t)|^2 = O(t^2) \rightarrow \infty \text{ with } t$$

No optimal solution (values unbounded).

Regularization in case $\mathbf{d} \notin \mathbf{R}(\mathbf{S})$

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T \mathbf{S} \mathbf{x} = 1\} \quad (\text{P})$$

Denote

$$Q = P_{N(S)} = I - S^\dagger S$$

$$\hat{S} = S + \kappa Q$$

$$\therefore \hat{S}^{-1} = S^\dagger + \frac{1}{\kappa} Q$$

and consider the problem

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T \hat{S} \mathbf{x} = 1\} \quad (\hat{\text{P}})$$

with optimal solution

$$\begin{aligned}\mathbf{x} &= \frac{1}{\sqrt{\mathbf{d}^T \widehat{S}^{-1} \mathbf{d}}} \widehat{S}^{-1} \mathbf{d} \\ &= \frac{1}{\sqrt{\mathbf{d}^T (S^\dagger + \frac{1}{\kappa} Q) \mathbf{d}}} \left(S^\dagger + \frac{1}{\kappa} Q \right) \mathbf{d}\end{aligned}$$

and optimal value

$$(\mathbf{d}^T \mathbf{x})^2 = \frac{A^2 + \frac{2AB}{\kappa} + \frac{B^2}{\kappa^2}}{A + \frac{B}{\kappa}}$$

where $A = (\mathbf{d}^T S^\dagger \mathbf{d})$, $B = \|z\|^2$

Two populations, equal covariance

The problem $\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\}$ (P)

where $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \notin R(S)$. Let $\hat{S} = S + \kappa Q$, $Q = P_{N(S)}$.

Two populations, equal covariance

The problem $\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\}$ (P)

where $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \notin R(S)$. Let $\hat{S} = S + \kappa Q$, $Q = P_{N(S)}$.

Then $\hat{S}^\dagger = S^\dagger + \frac{1}{\kappa} Q$ and the problem

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T \hat{S} \mathbf{x} = 1\} \quad (\hat{P})$$

Two populations, equal covariance

The problem $\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T S \mathbf{x} = 1\}$ (P)

where $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \notin R(S)$. Let $\hat{S} = S + \kappa Q$, $Q = P_{N(S)}$.
Then $\hat{S}^\dagger = S^\dagger + \frac{1}{\kappa} Q$ and the problem

$$\max \{(\mathbf{d}^T \mathbf{x})^2 : \mathbf{x}^T \hat{S} \mathbf{x} = 1\} \quad (\hat{P})$$

has the solution

$$\begin{aligned} \hat{\mathbf{x}} &= \frac{1}{\sqrt{\mathbf{d}^T \hat{S}^\dagger \mathbf{d}}} \hat{S}^\dagger \mathbf{d} \\ &= \frac{1}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d} + \frac{1}{\kappa} \|P_{N(S)} \mathbf{d}\|^2}} \left(S^\dagger \mathbf{d} + \frac{1}{\kappa} P_{N(S)} \mathbf{d} \right) \end{aligned}$$

$$\xrightarrow{\kappa \rightarrow \infty} \frac{1}{\sqrt{\mathbf{d}^T S^\dagger \mathbf{d}}} S^\dagger \mathbf{d}, \text{ the solution of (P)}$$

Two populations, equal covariance Σ

Let X_1, X_2 be the observations in \mathbb{R}^p from the two populations, and imbed in \mathbb{R}^{p+1} as follows:

Two populations, equal covariance Σ

Let X_1, X_2 be the observations in \mathbb{R}^p from the two populations, and imbed in \mathbb{R}^{p+1} as follows:

Points $\mathbf{x} \in X_1$ are shifted up to $z = 1$, i.e. $\mathbf{x} \rightarrow \hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$,

Two populations, equal covariance Σ

Let X_1, X_2 be the observations in \mathbb{R}^p from the two populations, and imbed in \mathbb{R}^{p+1} as follows:

Points $\mathbf{x} \in X_1$ are shifted up to $z = 1$, i.e. $\mathbf{x} \rightarrow \hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$,

points $\mathbf{x} \in X_2$ are shifted down to $z = -1$, i.e. $\mathbf{x} \rightarrow \hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ -1 \end{pmatrix}$.

Two populations, equal covariance Σ

Let X_1, X_2 be the observations in \mathbb{R}^p from the two populations, and imbed in \mathbb{R}^{p+1} as follows:

Points $\mathbf{x} \in X_1$ are shifted up to $z = 1$, i.e. $\mathbf{x} \rightarrow \hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$,

points $\mathbf{x} \in X_2$ are shifted down to $z = -1$, i.e. $\mathbf{x} \rightarrow \hat{\mathbf{x}} = \begin{pmatrix} \mathbf{x} \\ -1 \end{pmatrix}$.

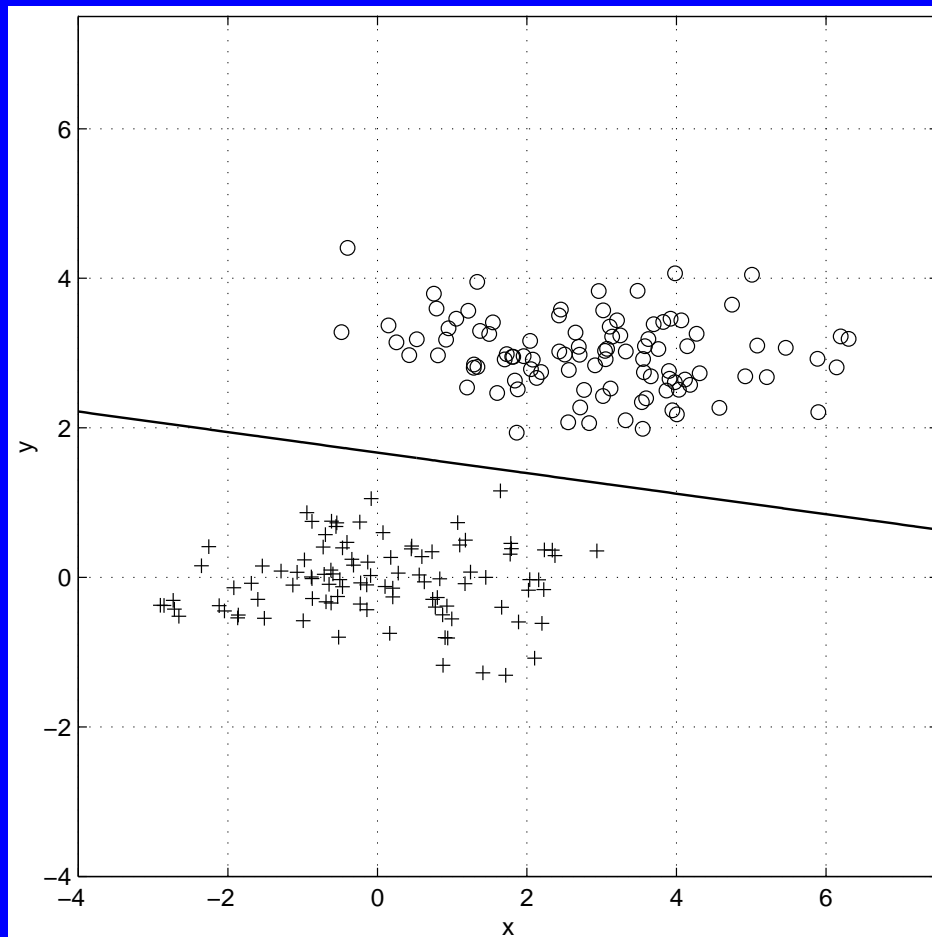
The covariance matrix

$$\hat{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$$

is singular even if Σ is not.

Two populations in \mathbb{R}^2

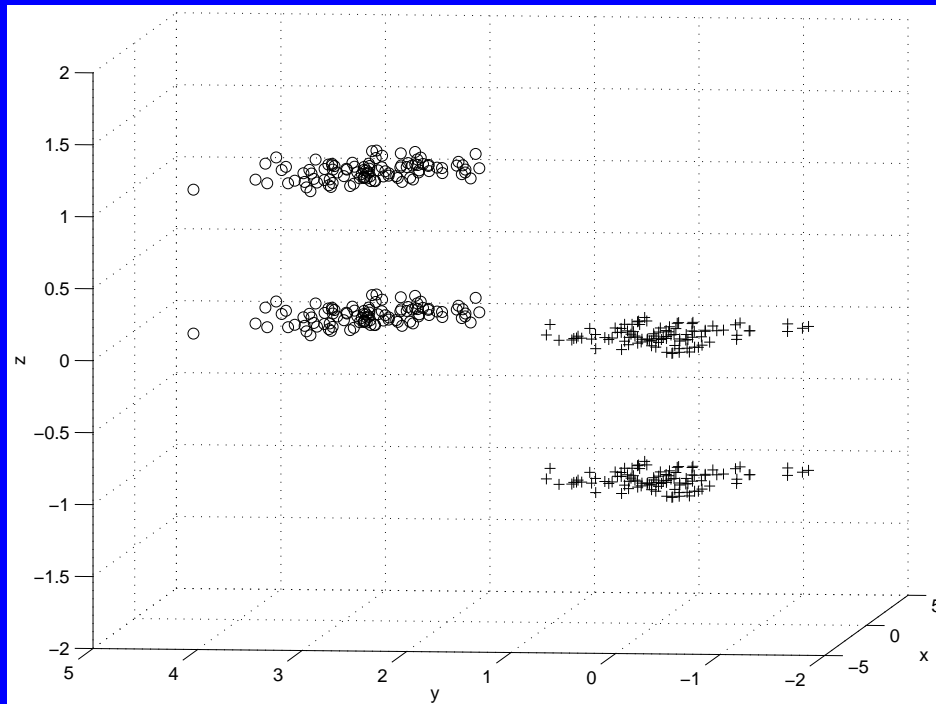
- $x \sim N(3, 1.5)$
- $y \sim N(3, 0.5)$
- + $x \sim N(0, 1.5)$
- + $y \sim N(0, 0.5)$



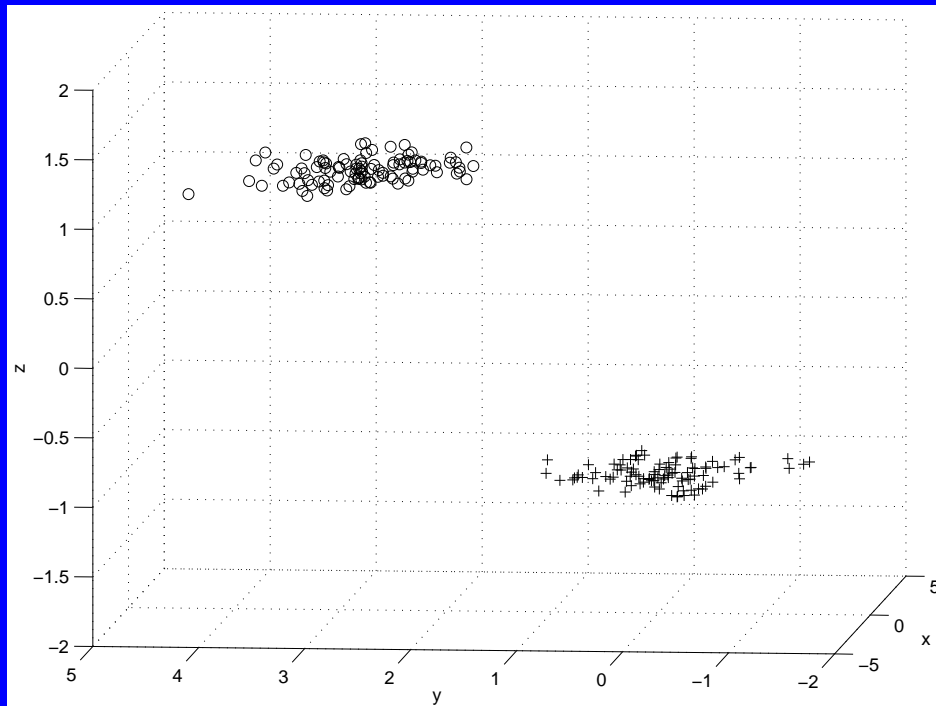
From \mathbb{R}^2 to \mathbb{R}^3

○ \longrightarrow $z = 1$

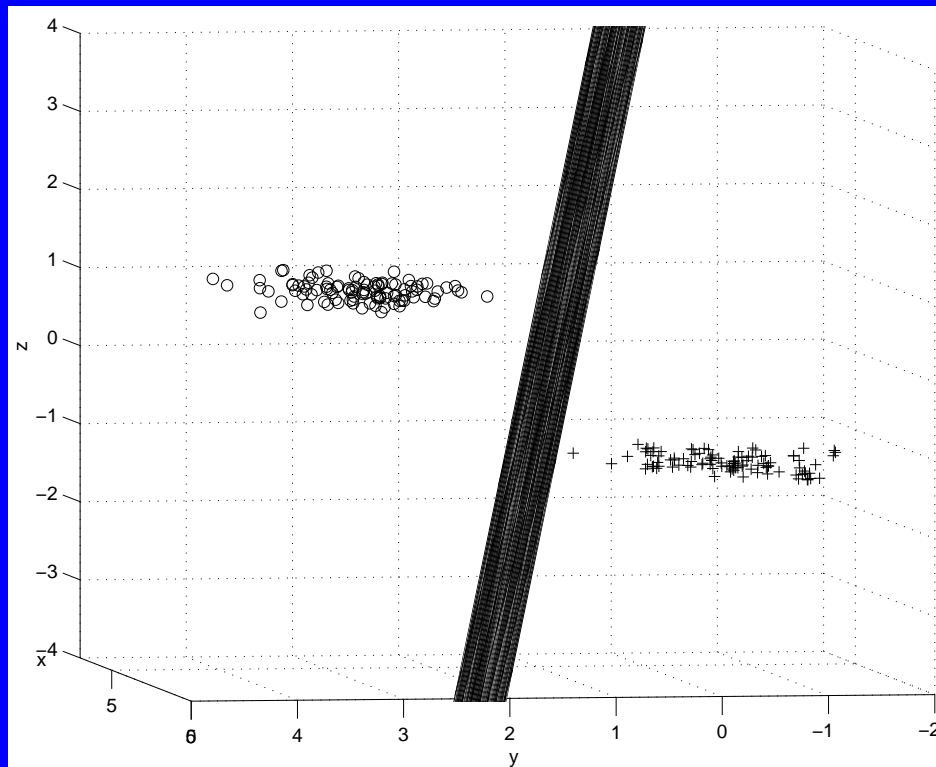
⊕ \longrightarrow $z = -1$



In \mathbb{R}^3

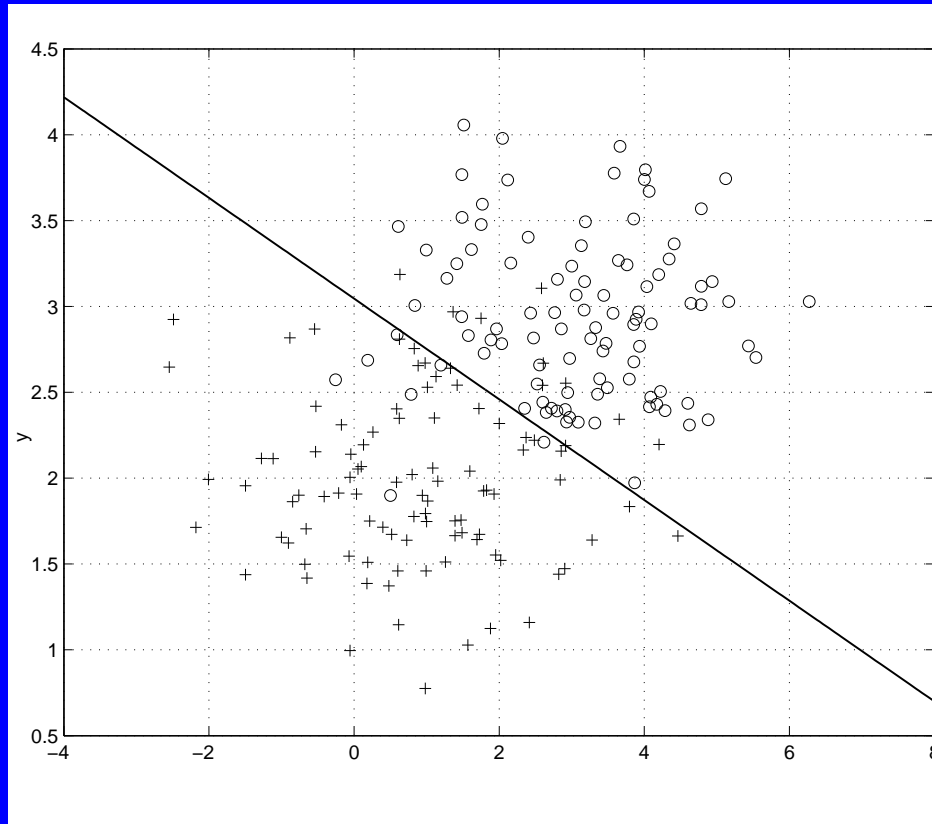


Separation in \mathbb{R}^3



Two populations in \mathbb{R}^2

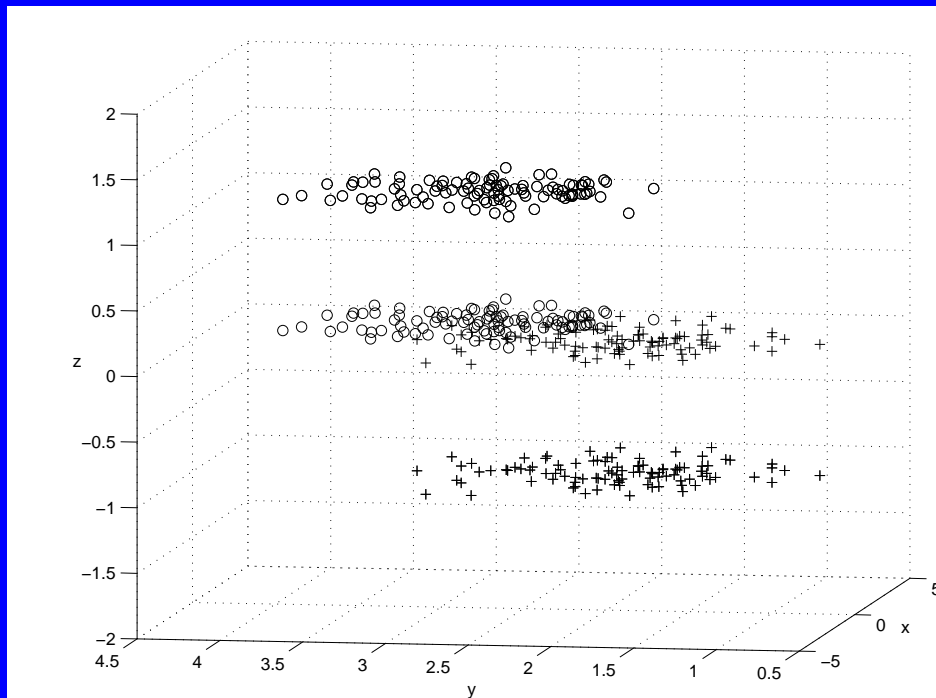
- $x \sim N(3, 1.5)$
- $y \sim N(3, 0.5)$
- + $x \sim N(1, 1.5)$
- + $y \sim N(2, 0.5)$



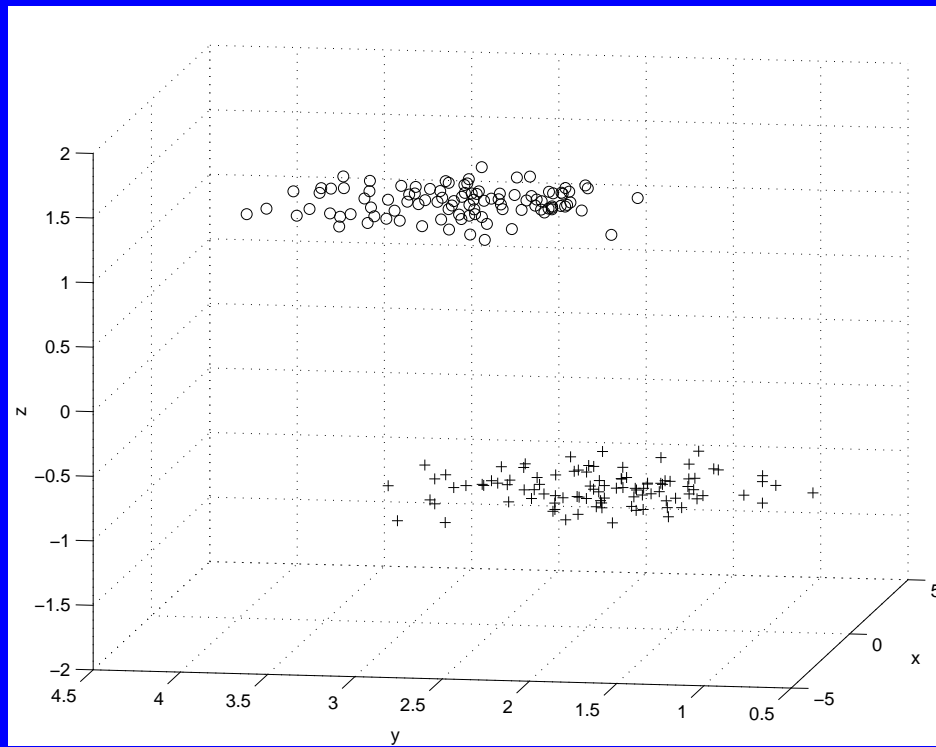
From \mathbb{R}^2 to \mathbb{R}^3

○ \longrightarrow $z = 1$

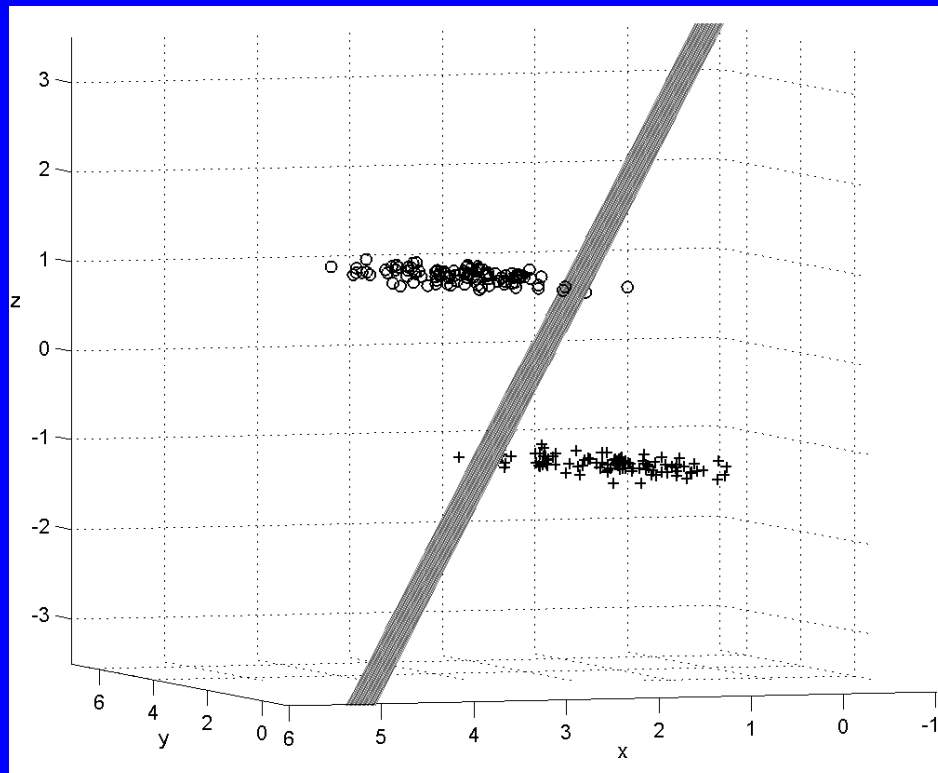
⊕ \longrightarrow $z = -1$



In \mathbb{R}^3

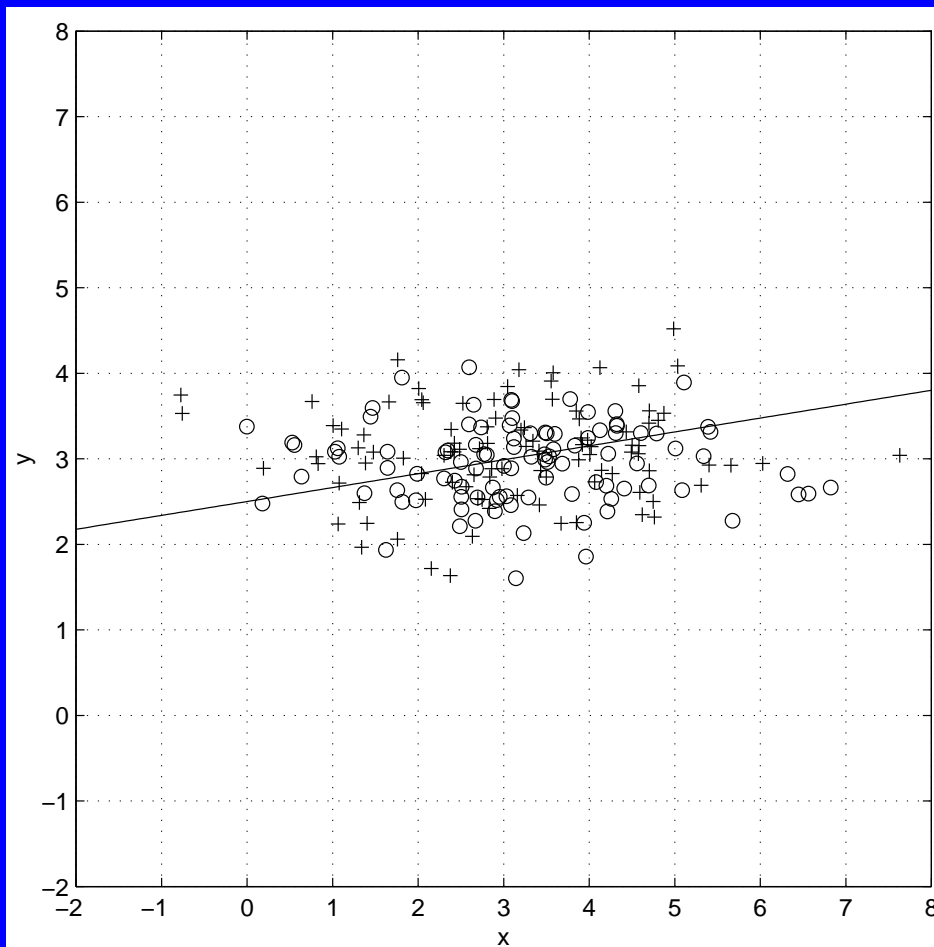


Separation in \mathbb{R}^3



Two populations in \mathbb{R}^2

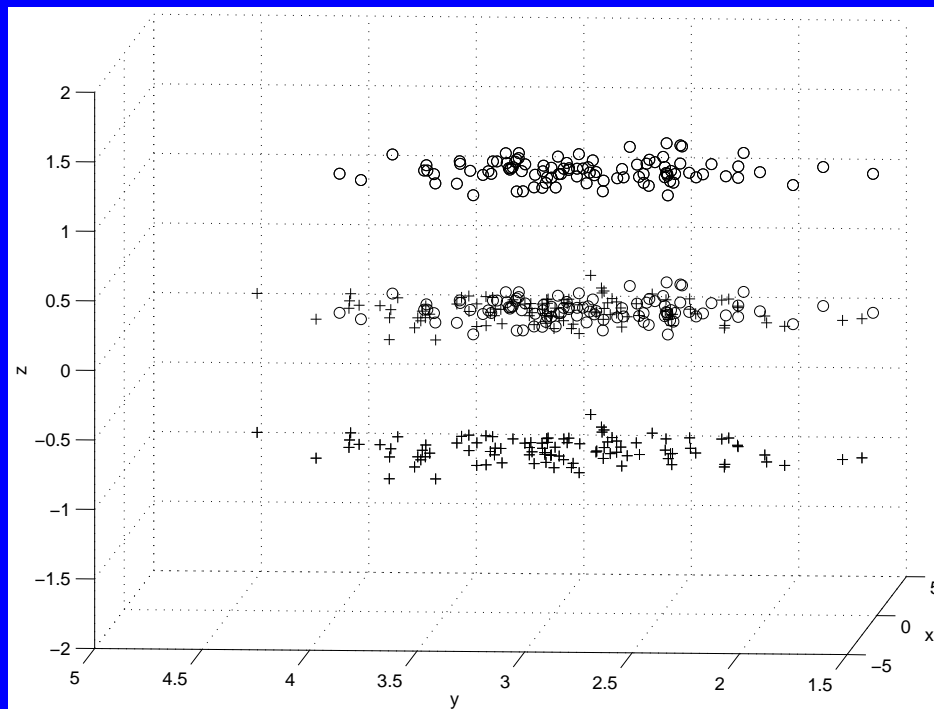
- $x \sim N(3, 1.5)$ + $x \sim N(3, 1.5)$
- $y \sim N(3, 0.5)$ + $y \sim N(3, 0.5)$



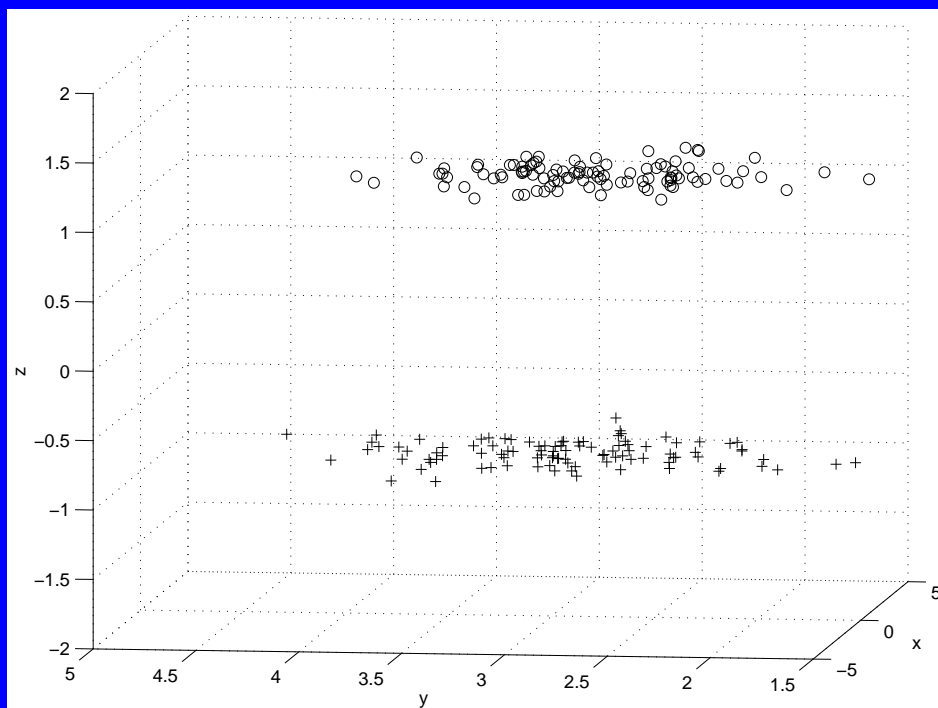
From \mathbb{R}^2 to \mathbb{R}^3

○ \longrightarrow $z = 1$

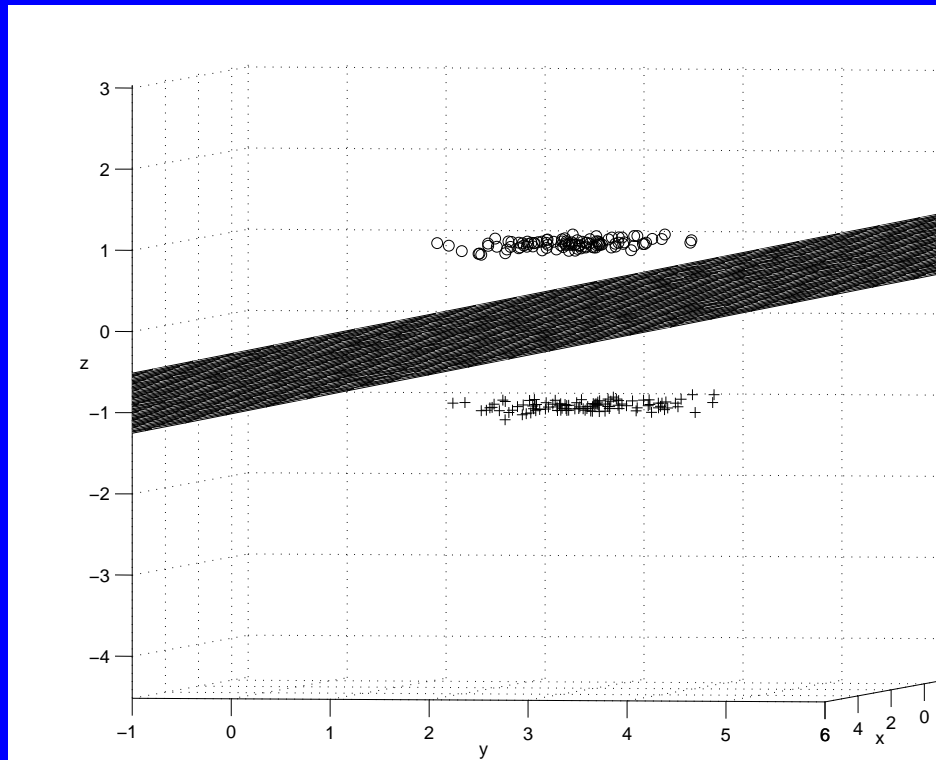
⊕ \longrightarrow $z = -1$



In \mathbb{R}^3



Separation in \mathbb{R}^3



Two populations, equal covariance Σ (contd.)

The problem

$$\max \{(\hat{\mathbf{d}}^T \mathbf{x})^2 : \mathbf{x}^T \hat{\mathbf{S}} \mathbf{x} = 1\} \quad (\hat{\mathbf{P}})$$

$$\text{where } \hat{\mathbf{d}} = \hat{\bar{\mathbf{x}}}_1 - \hat{\bar{\mathbf{x}}}_2 = \begin{pmatrix} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \\ 2 \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix},$$

Two populations, equal covariance Σ (contd.)

The problem

$$\max \{(\hat{\mathbf{d}}^T \mathbf{x})^2 : \mathbf{x}^T \hat{S} \mathbf{x} = 1\} \quad (\hat{P})$$

where $\hat{\mathbf{d}} = \hat{\bar{\mathbf{x}}}_1 - \hat{\bar{\mathbf{x}}}_2 = \begin{pmatrix} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \\ 2 \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix}$, has solution

$$\hat{\mathbf{x}} \propto \hat{S}^\dagger \hat{\mathbf{d}} = \begin{pmatrix} S^\dagger & 0 \\ 0 & \frac{1}{\kappa} \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix} = \begin{pmatrix} S^\dagger \mathbf{d} \\ \frac{2}{\kappa} \end{pmatrix}$$

Two populations, equal covariance Σ (contd.)

The problem

$$\max \{(\hat{\mathbf{d}}^T \mathbf{x})^2 : \mathbf{x}^T \hat{S} \mathbf{x} = 1\} \quad (\hat{P})$$

where $\hat{\mathbf{d}} = \hat{\bar{\mathbf{x}}}_1 - \hat{\bar{\mathbf{x}}}_2 = \begin{pmatrix} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \\ 2 \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix}$, has solution

$$\hat{\mathbf{x}} \propto \hat{S}^\dagger \hat{\mathbf{d}} = \begin{pmatrix} S^\dagger & 0 \\ 0 & \frac{1}{\kappa} \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix} = \begin{pmatrix} S^\dagger \mathbf{d} \\ \frac{2}{\kappa} \end{pmatrix}$$

It is the normal of the hyperplane separating \hat{X}_1, \hat{X}_2 in \mathbb{R}^{p+1} .

Two populations, equal covariance Σ (contd.)

The problem

$$\max \{(\hat{\mathbf{d}}^T \mathbf{x})^2 : \mathbf{x}^T \hat{S} \mathbf{x} = 1\} \quad (\hat{P})$$

where $\hat{\mathbf{d}} = \hat{\bar{\mathbf{x}}}_1 - \hat{\bar{\mathbf{x}}}_2 = \begin{pmatrix} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \\ 2 \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix}$, has solution

$$\hat{\mathbf{x}} \propto \hat{S}^\dagger \hat{\mathbf{d}} = \begin{pmatrix} S^\dagger & 0 \\ 0 & \frac{1}{\kappa} \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ 2 \end{pmatrix} = \begin{pmatrix} S^\dagger \mathbf{d} \\ \frac{2}{\kappa} \end{pmatrix}$$

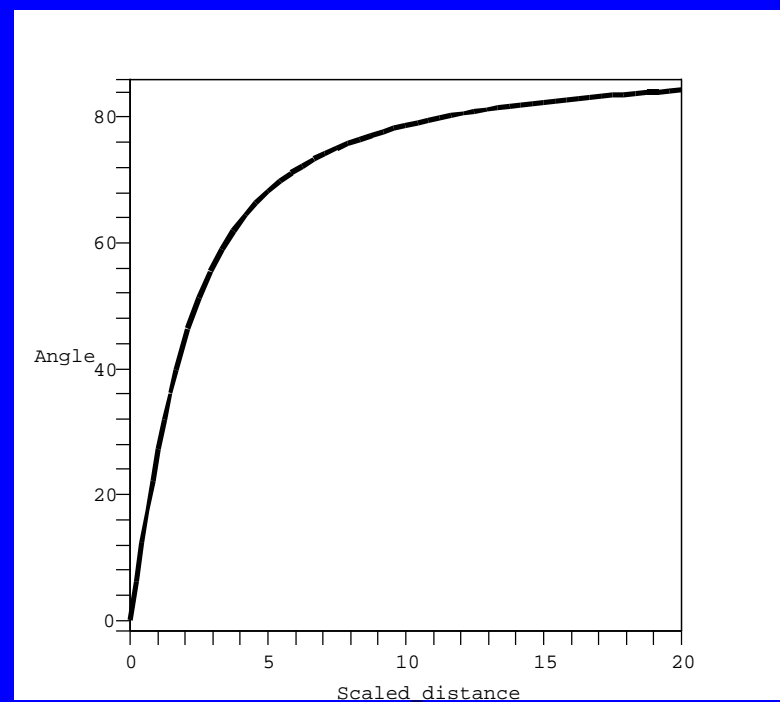
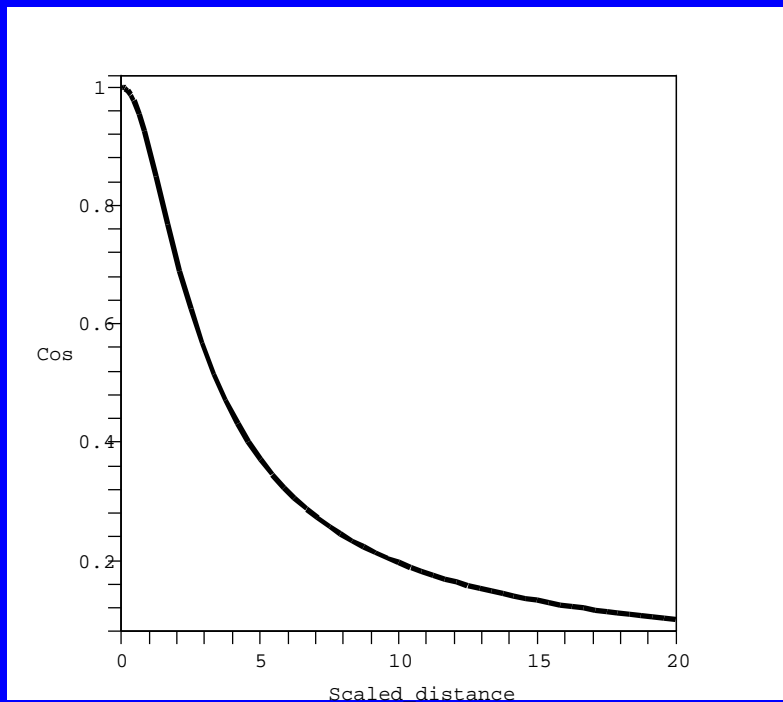
It is the normal of the hyperplane separating \hat{X}_1, \hat{X}_2 in \mathbb{R}^{p+1} .

The angle θ between this vector and the z -axis is given by

$$\cos \theta = \frac{\frac{2}{\kappa}}{\sqrt{\|S^\dagger \mathbf{d}\|^2 + \frac{4}{\kappa^2}}}$$

Angle of Separation

$$\theta = \arccos \frac{\frac{2}{\kappa}}{\sqrt{\|S^\dagger \mathbf{d}\|^2 + \frac{4}{\kappa^2}}} = \arctan \frac{\kappa \|S^\dagger \mathbf{d}\|}{2}$$



Angle of separation as a function of the scaled distance $\|S^\dagger \mathbf{d}\|$

Angle of separation θ for 5 datasets

Name of Data Set	$\cos \theta$	θ	% Correct
Breast Cancer	0.74	43°	96.5
Liver	0.99	4°	63.2
Diabetes	0.99	3°	74.7
Voting	0.18	80°	92.0
Hepatitis	0.42	65°	86.0

A Decomposition of Mahalanobis Distance

The Mahalanobis distance of $\boldsymbol{\mu} \in \mathbb{R}^q$ from $\mathbf{0}$ is

$$\Delta_q^2 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

A Decomposition of Mahalanobis Distance

The Mahalanobis distance of $\boldsymbol{\mu} \in \mathbb{R}^q$ from $\mathbf{0}$ is

$$\Delta_q^2 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

Partition $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)$, $\boldsymbol{\mu}_1 \in \mathbb{R}^k$ and correspondingly

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

A Decomposition of Mahalanobis Distance

The Mahalanobis distance of $\boldsymbol{\mu} \in \mathbb{R}^q$ from $\mathbf{0}$ is

$$\Delta_q^2 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

Partition $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)$, $\boldsymbol{\mu}_1 \in \mathbb{R}^k$ and correspondingly

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\begin{aligned} \therefore \Delta_q^2 &= \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_{2,1}^T \boldsymbol{\Sigma}_{22,1}^{-1} \boldsymbol{\mu}_{2,1} \\ &= \Delta_k^2 + \boldsymbol{\mu}_{2,1}^T \boldsymbol{\Sigma}_{22,1}^{-1} \boldsymbol{\mu}_{2,1} \end{aligned}$$

A Decomposition of Mahalanobis Distance

The Mahalanobis distance of $\boldsymbol{\mu} \in \mathbb{R}^q$ from $\mathbf{0}$ is

$$\Delta_q^2 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

Partition $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)$, $\boldsymbol{\mu}_1 \in \mathbb{R}^k$ and correspondingly

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\begin{aligned} \therefore \Delta_q^2 &= \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_{2,1}^T \boldsymbol{\Sigma}_{22,1}^{-1} \boldsymbol{\mu}_{2,1} \\ &= \Delta_k^2 + \boldsymbol{\mu}_{2,1}^T \boldsymbol{\Sigma}_{22,1}^{-1} \boldsymbol{\mu}_{2,1} \end{aligned}$$

where $\boldsymbol{\mu}_{2,1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_{22,1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$

A Decomposition of Mahalanobis Distance

The Mahalanobis distance of $\boldsymbol{\mu} \in \mathbb{R}^q$ from $\mathbf{0}$ is

$$\Delta_q^2 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

Partition $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)$, $\boldsymbol{\mu}_1 \in \mathbb{R}^k$ and correspondingly

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\begin{aligned} \therefore \Delta_q^2 &= \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_{2,1}^T \boldsymbol{\Sigma}_{22,1}^{-1} \boldsymbol{\mu}_{2,1} \\ &= \Delta_k^2 + \boldsymbol{\mu}_{2,1}^T \boldsymbol{\Sigma}_{22,1}^{-1} \boldsymbol{\mu}_{2,1} \end{aligned}$$

where $\boldsymbol{\mu}_{2,1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_{22,1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$

K.V. Mardia, J.T. Kent and J.M. Bibby, *Multivariate Analysis*,
Academic Press, 1979

Decomposition of M.d. (contd.)

If $\mathbf{u} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $M \sim W_q(\boldsymbol{\Sigma}, m)$ then the sample M.d.

$$D_q^2 = m\mathbf{u}^T M^{-1} \mathbf{u}$$

Decomposition of M.d. (contd.)

If $\mathbf{u} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $M \sim W_q(\boldsymbol{\Sigma}, m)$ then the **sample M.d.**

can be partitioned as $D_q^2 = m\mathbf{u}^T M^{-1} \mathbf{u}$

$$D_q^2 = D_k^2 + m\mathbf{z}^T M_{22,1}^{-1} \mathbf{z}$$

Decomposition of M.d. (contd.)

If $\mathbf{u} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $M \sim W_q(\boldsymbol{\Sigma}, m)$ then the **sample M.d.**

can be partitioned as $D_q^2 = m\mathbf{u}^T M^{-1} \mathbf{u}$

$$D_q^2 = D_k^2 + m\mathbf{z}^T M_{22,1}^{-1} \mathbf{z}$$

where $D_k^2 = m\mathbf{u}_1^T M_{11}^{-1} \mathbf{u}_1$, $M_{22,1} = M_{22} - M_{21}M_{11}^{-1}M_{12}$,

$$\mathbf{z} = \mathbf{u}_2 - M_{21}M_{11}^{-1} \mathbf{u}_1$$

Decomposition of M.d. (contd.)

If $\mathbf{u} \sim N(\boldsymbol{\mu}, \Sigma)$, $M \sim W_q(\Sigma, m)$ then the **sample M.d.**

can be partitioned as $D_q^2 = m\mathbf{u}^T M^{-1} \mathbf{u}$

$$D_q^2 = D_k^2 + m\mathbf{z}^T M_{22,1}^{-1} \mathbf{z}$$

where $D_k^2 = m\mathbf{u}_1^T M_{11}^{-1} \mathbf{u}_1$, $M_{22,1} = M_{22} - M_{21}M_{11}^{-1}M_{12}$,

$$\mathbf{z} = \mathbf{u}_2 - M_{21}M_{11}^{-1} \mathbf{u}_1$$

Theorem. If D_q^2 and D_k^2 are as above and $\boldsymbol{\mu}_{2,1} = \mathbf{0}$ then

$$\frac{D_q^2 - D_k^2}{m + D_k^2} \sim \frac{q - k}{m - q + 1} F_{q-k, m-q+1}$$

and is independent of D_k^2 . (Mardia et al, Theorem 3.6.2)

Decomposition of M.d. (contd.)

Let X_1, X_2 be samples in \mathbb{R}^p , with n_1, n_2 observations resp., from two populations $\sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, and let

$$n = n_1 + n_2, \quad c = \frac{n}{n_1 n_2}.$$

Decomposition of M.d. (contd.)

Let X_1, X_2 be samples in \mathbb{R}^p , with n_1, n_2 observations resp., from two populations $\sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, and let

$$n = n_1 + n_2, \quad c = \frac{n}{n_1 n_2}.$$

Imbed in \mathbb{R}^{p+1} by associating points \mathbf{x} with $\widehat{\mathbf{x}} = (z, \mathbf{x})$ where

$$z \sim N(1, 1) \quad \text{for } \mathbf{x} \in X_1, \quad z \sim N(-1, 1) \quad \text{for } \mathbf{x} \in X_2.$$

Decomposition of M.d. (contd.)

Let X_1, X_2 be samples in \mathbb{R}^p , with n_1, n_2 observations resp., from two populations $\sim N_p(\boldsymbol{\mu}_i, \Sigma)$, $i = 1, 2$, and let

$$n = n_1 + n_2, \quad c = \frac{n}{n_1 n_2}.$$

Imbed in \mathbb{R}^{p+1} by associating points \mathbf{x} with $\widehat{\mathbf{x}} = (z, \mathbf{x})$ where

$$z \sim N(1, 1) \quad \text{for } \mathbf{x} \in X_1, \quad z \sim N(-1, 1) \quad \text{for } \mathbf{x} \in X_2.$$

Then $\widehat{\mathbf{d}} = \widehat{\bar{\mathbf{x}}}_1 - \widehat{\bar{\mathbf{x}}}_2 = (\bar{z}_1 - \bar{z}_2, \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = (\bar{z}_1 - \bar{z}_2, \bar{\mathbf{d}})$ has covariance matrix

$$\widehat{\Sigma} = c \begin{pmatrix} 1 & 0 \\ 0 & \Sigma \end{pmatrix}$$

Decomposition of M.d. (contd.)

Let X_1, X_2 be samples in \mathbb{R}^p , with n_1, n_2 observations resp., from two populations $\sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, and let

$$n = n_1 + n_2, \quad c = \frac{n}{n_1 n_2}.$$

Imbed in \mathbb{R}^{p+1} by associating points \mathbf{x} with $\widehat{\mathbf{x}} = (z, \mathbf{x})$ where

$$z \sim N(1, 1) \quad \text{for } \mathbf{x} \in X_1, \quad z \sim N(-1, 1) \quad \text{for } \mathbf{x} \in X_2.$$

Then $\widehat{\mathbf{d}} = \widehat{\mathbf{x}}_1 - \widehat{\mathbf{x}}_2 = (\bar{z}_1 - \bar{z}_2, \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = (\bar{z}_1 - \bar{z}_2, \bar{\mathbf{d}})$ has covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = c \begin{pmatrix} 1 & 0 \\ 0 & \boldsymbol{\Sigma} \end{pmatrix}$$

If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ then

$$\widehat{\mathbf{d}} \sim N_q((2, \mathbf{0}), \widehat{\boldsymbol{\Sigma}})$$

Decomposition of M.d. (contd.)

If S_1, S_2 are the variances of 2 samples, the **pooled variance** is

$$S_{\text{pooled}} = \frac{n_1 S_1 + n_2 S_2}{n - 2},$$

Decomposition of M.d. (contd.)

If S_1, S_2 are the variances of 2 samples, the **pooled variance** is

$$S_{\text{pooled}} = \frac{n_1 S_1 + n_2 S_2}{n - 2},$$

$$(n - 2) S_{\text{pooled}} \sim W_p(\Sigma, n - 2),$$

$$\begin{aligned} \text{and } c^{-1} \bar{\mathbf{d}}^T S_{\text{pooled}}^{-1} \bar{\mathbf{d}} &\sim T^2(p, n - 2) \\ &\sim \frac{(n - 2)p}{n - p - 1} F_{p, n - p - 1}. \end{aligned}$$

Decomposition of M.d. (contd.)

If S_1, S_2 are the variances of 2 samples, the **pooled variance** is

$$S_{\text{pooled}} = \frac{n_1 S_1 + n_2 S_2}{n - 2},$$

$$(n - 2) S_{\text{pooled}} \sim W_p(\Sigma, n - 2),$$

$$\begin{aligned} \text{and } c^{-1} \bar{\mathbf{d}}^T S_{\text{pooled}}^{-1} \bar{\mathbf{d}} &\sim T^2(p, n - 2) \\ &\sim \frac{(n - 2)p}{n - p - 1} F_{p, n - p - 1}. \end{aligned}$$

The Mahalanobis distance $D_{p+1}^2 = c^{-1} \bar{\mathbf{d}}^T \hat{S}^{-1} \bar{\mathbf{d}}$ is decomposed

$$\begin{aligned} D_{p+1}^2 &= D_1^2 + c^{-1} \bar{\mathbf{d}}^T \hat{S}_{22,1}^{-1} \bar{\mathbf{d}} \\ &= c^{-1} (\bar{z}_1 - \bar{z}_2)^2 + c^{-1} \bar{\mathbf{d}}^T S^{-1} \bar{\mathbf{d}} \end{aligned}$$

Decomposition of M.d. (contd.)

Theorem. If $\mu_1 = \mu_2$ then

$$\begin{aligned} \frac{D_{p+1}^2 - D_1^2}{(n-2) + D_1^2} &= \frac{c^{-1} \bar{\mathbf{d}}^T S^{-1} \bar{\mathbf{d}}}{(n-2) + c^{-1} (\bar{z}_1 - \bar{z}_2)^2} \\ &\sim \frac{p}{n-2-p} F_{p, n-2-p} \end{aligned}$$

Decomposition of M.d. (contd.)

Theorem. If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ then

$$\begin{aligned} \frac{D_{p+1}^2 - D_1^2}{(n-2) + D_1^2} &= \frac{c^{-1} \bar{\mathbf{d}}^T S^{-1} \bar{\mathbf{d}}}{(n-2) + c^{-1} (\bar{z}_1 - \bar{z}_2)^2} \\ &\sim \frac{p}{n-2-p} F_{p, n-2-p} \end{aligned}$$

Corollary. Let $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and let θ be the angle of separation for the normalized observations $\Sigma^{-1/2} \mathbf{x}$. Then

$$\tan^2 \theta \sim \frac{p}{n-2-p} F_{p, n-2-p}$$

