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 $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^{p+1}$

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

$y \in Y \subset \mathbb{R}$ (output, class) more difficult to measure.
The objects of study are vectors $v = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^{p+1}$

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

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

**Problem**: Predict (or estimate) $y$ given $x$.

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

$$D = \{(x_i, y_i) : i = 1, \ldots, N\}$$
Statistical Learning

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

\( x \in \mathbf{X} \subset \mathbb{R}^{p} \) (inputs, attributes) observable, readily measurable.
\( y \in \mathbf{Y} \subset \mathbb{R} \) (output, class) more difficult to measure.

**Problem:** Predict (or estimate) \( y \) given \( x \).

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

\[
\mathbf{D} = \{(x_i, y_i) : i = 1, \ldots, N\}
\]

**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

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

- $x \sim N(3, 1.5) \quad + \quad x \sim N(1, 1.5)$
- $y \sim N(3, 0.5) \quad + \quad y \sim N(2, 0.5)$
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: $\mathbf{x} = (x_1, \ldots, x_p)$ results of diagnostic tests, $y \in \{0, 1\}$ denoting respectively the absence or presence of disease.
Medical applications

Typically: \( \mathbf{x} = (x_1, \ldots, x_p) \) 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: \( x = (x_1, \ldots, x_p) \) 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: \textbf{false positive}, and
- type 2: \textbf{false negative}, differ in their consequences.

\url{www.ics.uci.edu/~mlearn/MLRepository.html}
Medical applications

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

\[ y \text{ dictates the course of treatment.} \]

The two possible errors:
- type 1: false positive, and
- type 2: false negative, differ in their consequences.


A Naive Proposal

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

$$d\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(x_1, x_2) + \alpha d_Y^2(y_1, y_2)}$$
A Naive Proposal

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

$$d\left( \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \right) = \sqrt{d_X^2(x_1, x_2) + \alpha d_Y^2(y_1, y_2)}$$

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

A Naive Proposal

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

\[
d\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d^2_X(x_1, x_2) + \alpha d^2_Y(y_1, y_2)}
\]

2. Use \( d \) for classification of \( \mathcal{D} \) in clusters \( \{ \Omega_1, \ldots, \Omega_m \} \).

3. For each cluster \( \Omega_i \) compute:
   - a center \( \overline{y}_i \) of \( y \),
   - \( \Omega_i^X \) the \( X \)-projection of \( \Omega_i \)
A Naive Proposal

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

$$d\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(x_1, x_2) + \alpha d_Y^2(y_1, y_2)}$$

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

3. For each cluster $\Omega_i$ compute:
   - a center $\overline{y}_i$ of $y$,
   - $\Omega_i^X$ the $X$–projection of $\Omega_i$

4. Given $\mathbf{x} \in X$, determine the nearest projected cluster, say $\Omega_i^X$. 
A Naive Proposal

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

$$d\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d_X^2(x_1, x_2) + \alpha d_Y^2(y_1, y_2)}$$

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

3. For each cluster $\Omega_i$ compute:
   a center $\overline{y}_i$ of $y$,
   $\Omega^X_i$ the $X$-projection of $\Omega_i$

4. Given $x \in X$, determine the nearest projected cluster, say $\Omega^X_i$.

5. Use $\overline{y}_i$ as estimate for $y$. 
A Naive Proposal

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

$$d\left(\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}\right) = \sqrt{d^2_X(x_1, x_2) + \alpha d^2_Y(y_1, y_2)}$$

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

3. For each cluster $\Omega_i$ compute:
   a **center** $\overline{y}_i$ of $y$,
   $\Omega_i^X$ the $X$–projection of $\Omega_i$

4. Given $x \in X$, determine the nearest projected cluster, say $\Omega_i^X$.

5. Use $\overline{y}_i$ as estimate for $y$.

Yuri Levin and A. B–I, *Opsearch*, 2000
<table>
<thead>
<tr>
<th>Name of Data Set</th>
<th>% Correct Predictions</th>
<th>% Errors</th>
<th>Lim et al</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>96.5</td>
<td>100</td>
<td>93.1</td>
</tr>
<tr>
<td>Liver</td>
<td>63.2</td>
<td>79.3</td>
<td>49.7</td>
</tr>
<tr>
<td>Diabetes</td>
<td>74.7</td>
<td>79.9</td>
<td>65.7</td>
</tr>
<tr>
<td>Voting</td>
<td>92.0</td>
<td>98.78</td>
<td>82.3</td>
</tr>
<tr>
<td>Wine</td>
<td>93.7</td>
<td>100</td>
<td>82.35</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>86.03</td>
<td>96.42</td>
<td>71.43</td>
</tr>
</tbody>
</table>
Fisher's Discriminant: Separation of Populations with equal Covariances

Observations $\mathbf{x} \in \mathbb{R}^p$ from two populations with equal covariance $\Sigma$. 
Fisher’s Discriminant: Separation of Populations with equal Covariances

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

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

\[
\frac{(a^T \bar{x}_1 - a^T \bar{x}_2)^2}{a^T S a}
\]
Fisher’s Discriminant: Separation of Populations with equal Covariances

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

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

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

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

$$\frac{(\overline{y}_1 - \overline{y}_2)^2}{s_y^2} = \frac{(\mathbf{a}^T \overline{x}_1 - \mathbf{a}^T \overline{x}_2)^2}{\mathbf{a}^T S \mathbf{a}}.$$
Fisher’s Discriminant: Separation of Populations with equal Covariances

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

\[
\max \ \{(a^T d)^2 : a^T S a = 1\} \quad (P)
\]
**Fisher’s Discriminant: Separation of Populations with equal Covariances**

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

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

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 \( \mathbf{d} := \bar{x}_1 - \bar{x}_2 \). The problem:

\[
\max \left\{ (\mathbf{a}^T \mathbf{d})^2 : \mathbf{a}^T S \mathbf{a} = 1 \right\} \tag{P}
\]

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

The samples represented by ellipses have means $\bar{x}_i$, $i = 1, 2$ and variance $S$. 
\[ d = x_1 - x_2 \]
\[ d = x_1 - x_2 \]
\[ d = x_1 - x_2 \]

\[
\begin{align*}
\max & \quad \frac{(a^T d)^2}{a^T S a} \\
\text{subject to} & \quad (a^T d)^2 : a^T S a = 1
\end{align*}
\]
The Fisher discriminant is given by the line \( d^T S^{-1} x = \alpha \)
The Fisher discriminant is given by the line $d^T S^{-1} x = \alpha$

$$\alpha = \frac{1}{2} d^T S^{-1} (\bar{x}_1 + \bar{x}_2)$$
Classification using Fisher’s Discriminant
Classification using Fisher’s Discriminant

Let $\bar{x}_1$, $\bar{x}_2$, $d$, $S$ be as above.
Assign an observation $x$ to population 1 if

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

to population 2 otherwise.
Classification using Fisher’s Discriminant

Let $\bar{x}_1$, $\bar{x}_2$, $d$, $S$ be as above.
Assign an observation $x$ to population 1 if
\[ d^T S^{-1} x > \frac{1}{2} d^T S^{-1} (\bar{x}_1 + \bar{x}_2) \]
to population 2 otherwise.
An Optimization Problem

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

The problem:

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

(P)
An Optimization Problem

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

The problem:

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

Lagrangian:

$$L(x, \lambda) = (d^T x)^2 - \lambda (x^T S x - 1)$$
An Optimization Problem

d ∈ ℝⁿ, S ∈ ℝⁿ×ⁿ PSD.

The problem:

\[
\max \{ (d^T x)^2 : x^T S x = 1 \} \tag{P}
\]

Lagrangian:

\[
L(x, \lambda) = (d^T x)^2 - \lambda (x^T S x - 1)
\]

An optimal solution must satisfy

\[
\frac{1}{2} \nabla L(x, \lambda) = (d^T x) d - \lambda S x = 0
\]
An Optimization Problem

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

The problem:

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

Lagrangian:

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

An optimal solution must satisfy

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

$$\therefore S x = \left( \frac{d^T x}{\lambda} \right) d \quad \text{(1)}$$
Case 1: \( d \in R(S) \)

\[
x = \left( \frac{d^T x}{\lambda} \right) S^\dagger d
\]

(2)

\[
\therefore x = \alpha S^\dagger d, \quad \alpha = \frac{d^T x}{\lambda}
\]

(3)

\[
\therefore x^T S x = \alpha^2 d^T S^\dagger S S^\dagger d = \alpha^2 d^T S^\dagger d = 1
\]

(4)

\[
\therefore \alpha^2 = \frac{1}{d^T S^\dagger d}
\]

(4)

\[
\therefore x = \frac{1}{\sqrt{d^T S^\dagger d}} S^\dagger d
\]

(5)
An Optimization Problem (cont’d)

\[
\max \{(d^T x)^2 : x^T S x = 1\} \quad (P)
\]

The story so far: If \( d \in R(S) \) then

\[
x = \left( \frac{1}{\sqrt{d^T S^+ d}} \right) S^+ d \quad (5)
\]

\[
(d^T x)^2 = \left( \frac{d^T S^+ d}{\sqrt{d^T S^+ d}} \right)^2 = d^T S^+ d \quad (6)
\]

Case 2: \( d \notin R(S) \) (so \( S \) is singular)

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

Let \( x_0 \) satisfy

\[
x_0^T S x_0 = 1
\]

\[
x(t) := x_0 + tz
\]

\( \therefore x(t)^T S x(t) = 1, \ \forall \ t \)
An Optimization Problem (cont’d)

But

\[ d^T x(t) = d^T x_0 + t d^T z \]
\[ = d^T x_0 + t d^T P_{N(S)} d \]
\[ = d^T x_0 + t \| P_{N(S)} d \|^2 \]
\[ = d^T x_0 + t \| z \|^2 \]

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

No optimal solution (values unbounded).
Regularization in case $d \notin \mathbb{R}(S)$

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

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

$$\max \left\{ (d^T x)^2 : x^T \hat{S} x = 1 \right\} \quad (\hat{P})$$
with optimal solution

\[ x = \frac{1}{\sqrt{d^T \hat{S}^{-1} d}} \hat{S}^{-1} d \]

\[ = \frac{1}{\sqrt{d^T (S^\dagger + \frac{1}{\kappa} Q) d}} \left( S^\dagger + \frac{1}{\kappa} Q \right) d \]

and optimal value

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

where \( A = (d^T S^\dagger d) \), \( B = \|z\|^2 \)
Two populations, equal covariance

The problem

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

where $d = \bar{x}_1 - \bar{x}_2 \not\in R(S)$. Let $\hat{S} = S + \kappa Q$, $Q = P_{N(S)}$. 
Two populations, equal covariance

The problem
\[
\max \left\{ (d^T x)^2 : x^T S x = 1 \right\} \quad \text{(P)}
\]

where \( d = \bar{x}_1 - \bar{x}_2 \not\in R(S) \). Let \( \hat{S} = S + \kappa Q, \) \( Q = P_{N(S)} \).

Then \( \hat{S}^\dagger = S^\dagger + \frac{1}{\kappa} Q \) and the problem
\[
\max \left\{ (d^T x)^2 : x^T \hat{S} x = 1 \right\} \quad \text{(P)}
\]
Two populations, equal covariance

The problem

\[
\max \left\{ (d^T x)^2 : x^T S x = 1 \right\} \quad (P)
\]

where \( d = \bar{x}_1 - \bar{x}_2 \not\in R(S) \). Let \( \hat{S} = S + \kappa Q, Q = P_{N(S)} \). Then \( \hat{S}^\dagger = S^\dagger + \frac{1}{\kappa} Q \) and the problem

\[
\max \left\{ (d^T x)^2 : x^T \hat{S} x = 1 \right\} \quad (\hat{P})
\]

has the solution

\[
\hat{x} = \frac{1}{\sqrt{d^T \hat{S}^\dagger d}} \hat{S}^\dagger d
\]

\[
= \frac{1}{\sqrt{d^T S^\dagger d + \frac{1}{\kappa} \|P_{N(S)} d\|^2}} \left( S^\dagger d + \frac{1}{\kappa} P_{N(S)} d \right)
\]

as \( \kappa \to \infty \)

\[
\frac{1}{\sqrt{d^T S^\dagger d}} S^\dagger d, \text{ the solution of (P)}
\]
Two populations, equal covariance $\Sigma$

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 $\Sigma$

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 $x \in X_1$ are shifted up to $z = 1$, i.e. $x \rightarrow \hat{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$,
Two populations, equal covariance $\Sigma$

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

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

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

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

The covariance matrix

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

is singular even if $\Sigma$ is not.
Two populations in $\mathbb{R}^2$

- $x \sim N(3, 1.5)$  +  $x \sim N(0, 1.5)$
- $y \sim N(3, 0.5)$  +  $y \sim N(0, 0.5)$
From $\mathbb{R}^2$ to $\mathbb{R}^3$

- $\circ \quad \rightarrow \quad z = 1$
- $+ \quad \rightarrow \quad z = -1$
In $\mathbb{R}^3$
Separation in $\mathbb{R}^3$
Two populations in $\mathbb{R}^2$

- $x \sim N(3, 1.5)$  +  $x \sim N(1, 1.5)$
- $y \sim N(3, 0.5)$  +  $y \sim N(2, 0.5)$
From $\mathbb{R}^2$ to $\mathbb{R}^3$

$\circ \quad \rightarrow \quad z = 1$

$+ \quad \rightarrow \quad z = -1$
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 $\mathbb{R}^3$

- $\circ \quad \rightarrow \quad z = 1$
- $+ \quad \rightarrow \quad z = -1$
In $\mathbb{R}^3$
Separation in $\mathbb{R}^3$
Two populations, equal covariance $\Sigma$ (contd.)

The problem

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

where $\hat{d} = \hat{x}_1 - \hat{x}_2 = \begin{pmatrix} x_1 - x_2 \\ 2 \end{pmatrix}$, $\hat{x}_1$, $\hat{x}_2$
Two populations, equal covariance $\Sigma$ (contd.)

The problem

$$\max \left\{ (d^T x)^2 : x^T S x = 1 \right\}$$

$(\hat{P})$

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

$$\hat{x} \propto \hat{S}^\dagger \hat{d} = \begin{pmatrix} S^\dagger & 0 \\ 0 & \frac{1}{\kappa} \end{pmatrix} \begin{pmatrix} d \\ 2 \end{pmatrix} = \begin{pmatrix} S^\dagger d \\ \frac{2}{\kappa} \end{pmatrix}$$
Two populations, equal covariance $\Sigma$ (contd.)

The problem

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

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

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

It is the normal of the hyperplane separating $\hat{X}_1, \hat{X}_2$ in $\mathbb{R}^{p+1}$. 
Two populations, equal covariance $\Sigma$ (contd.)

The problem

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

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

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

It is the normal of the hyperplane separating $\hat{X}_1, \hat{X}_2$ in $\mathbb{R}^{p+1}$. The angle $\theta$ between this vector and the z-axis is given by

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

\[ \theta = \arccos \frac{2}{\kappa} \frac{\|S^\dagger d\|^2}{\sqrt{\|S^\dagger d\|^2 + 4/\kappa^2}} = \arctan \frac{\kappa \|S^\dagger d\|}{2} \]

Angle of separation as a function of the scaled distance $\|S^\dagger d\|$
Angle of separation $\theta$ for 5 datasets

<table>
<thead>
<tr>
<th>Name of Data Set</th>
<th>$\cos \theta$</th>
<th>$\theta$</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>0.74</td>
<td>43°</td>
<td>96.5</td>
</tr>
<tr>
<td>Liver</td>
<td>0.99</td>
<td>4°</td>
<td>63.2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>0.99</td>
<td>3°</td>
<td>74.7</td>
</tr>
<tr>
<td>Voting</td>
<td>0.18</td>
<td>80°</td>
<td>92.0</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>0.42</td>
<td>65°</td>
<td>86.0</td>
</tr>
</tbody>
</table>
A Decomposition of Mahalanobis Distance

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

$$\Delta_q^2 = \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}$$
A Decomposition of Mahalanobis Distance

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

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

Partition $\mu^T = (\mu_1^T, \mu_2^T)$, $\mu_1 \in \mathbb{R}^k$ and correspondingly

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$
A Decomposition of Mahalanobis Distance

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

$$\Delta^2_q = \mu^T \Sigma^{-1} \mu$$

Partition $\mu^T = (\mu_1^T, \mu_2^T)$, $\mu_1 \in \mathbb{R}^k$ and correspondingly

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

$$\therefore \Delta^2_q = \mu_1^T \Sigma_{11}^{-1} \mu_1 + \mu_2^T \Sigma_{22,1}^{-1} \mu_2,1$$

$$= \Delta^2_k + \mu_2^T \Sigma_{22,1}^{-1} \mu_2,1$$
A Decomposition of Mahalanobis Distance

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

$$\Delta^2_q = \mu^T \Sigma^{-1} \mu$$

Partition $\mu^T = (\mu_1^T, \mu_2^T)$, $\mu_1 \in \mathbb{R}^k$ and correspondingly

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

$$\therefore \Delta^2_q = \mu_1^T \Sigma_{11}^{-1} \mu_1 + \mu_2^T \Sigma_{22,1}^{-1} \mu_{2,1}$$

$$= \Delta^2_k + \mu_{2,1}^T \Sigma_{22,1}^{-1} \mu_{2,1}$$

where $\mu_{2,1} = \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} \mu_1$, $\Sigma_{22,1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$
A Decomposition of Mahalanobis Distance

The Mahalanobis distance of \( \mu \in \mathbb{R}^q \) from 0 is

\[
\Delta_q^2 = \mu^T \Sigma^{-1} \mu
\]

Partition \( \mu^T = (\mu_1^T, \mu_2^T) \), \( \mu_1 \in \mathbb{R}^k \) and correspondingly

\[
\Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix}
\]

\[
\therefore \Delta_q^2 = \mu_1^T \Sigma_{11}^{-1} \mu_1 + \mu_2^T \Sigma_{22,1}^{-1} \mu_{2,1}
\]

\[
= \Delta_k^2 + \mu_2^T \Sigma_{22,1}^{-1} \mu_{2,1}
\]

where \( \mu_{2,1} = \mu_2 - \Sigma_{21} \Sigma_{11}^{-1} \mu_1, \quad \Sigma_{22,1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \)

Decomposition of M.d. (contd.)

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

$$D_q^2 = m u^T M^{-1} u$$
Decomposition of M.d. (contd.)

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

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

$$D_q^2 = D_k^2 + m\mathbf{z}^T M_{22}^{-1} \mathbf{z}$$
Decomposition of M.d. (contd.)

If \( u \sim N(\mu, \Sigma) \), \( M \sim W_q(\Sigma, m) \) then the sample M.d.
can be partitioned as

\[ D_q^2 = m u^T M^{-1} u \]

\[ D_q^2 = D_k^2 + m z^T M_{22,1} z \]

where

\[ D_k^2 = mu_1^T M_{11}^{-1} u_1, \quad M_{22,1} = M_{22} - M_{21} M_{11}^{-1} M_{12}, \]

\[ z = u_2 - M_{21} M_{11}^{-1} u_1 \]
Decomposition of M.d. (contd.)

If \( u \sim N(\mu, \Sigma) \), \( M \sim W_q(\Sigma, m) \) then the sample M.d.
can be partitioned as

\[
D_q^2 = m u^T M^{-1} u
\]

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

where \( D_k^2 = m u_1^T M_{11}^{-1} u_1, \ M_{22,1} = M_{22} - M_{21} M_{11}^{-1} M_{12}, \)
\[
z = u_2 - M_{21} M_{11}^{-1} u_1
\]

Theorem. If \( D_q^2 \) and \( D_k^2 \) are as above and \( \mu_{2,1} = 0 \) 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 $\sim N_p(\mu_i, \Sigma)$, $i = 1, 2$, and let

$$n = n_1 + n_2, \quad c = \frac{n}{n_1n_2}.$$
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 \( \sim N_p(\mu_i, \Sigma) \), \( i = 1, 2 \), and let

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

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

\[
z \sim N(1, 1) \quad \text{for} \quad x \in X_1, \quad z \sim N(-1, 1) \quad \text{for} \quad x \in X_2.
\]
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 $\sim N_p(\mu_i, \Sigma)$, $i = 1, 2$, and let

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

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

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

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

$$\hat{\Sigma} = c \begin{pmatrix} 1 & 0 \\ 0 & \Sigma \end{pmatrix}.$$
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 $\sim N_p(\mu_i, \Sigma), \ i = 1, 2$, and let

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

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

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

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

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

If $\mu_1 = \mu_2$ then

$$\bar{d} \sim N_q((2, 0), \hat{\Sigma}).$$
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

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

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

and

$$c^{-1} \bar{d}^T S_{\text{pooled}}^{-1} \bar{d} \sim T^2(p, n - 2) \sim \frac{(n - 2)p}{n - p - 1} F_{p,n-p-1}.$$
Decomposition of M.d. (contd.)

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

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

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

and

\[c^{-1} \overline{d}^T S_{\text{pooled}}^{-1} \overline{d} \sim T^2(p, n - 2)\]

\[\sim \frac{(n - 2)p}{n - p - 1} F_{p,n-p-1}.\]

The Mahalanobis distance \( D_{p+1}^2 = c^{-1} \overline{d}^T \hat{S}^{-1} \overline{d} \) is decomposed

\[
D_{p+1}^2 = D_1^2 + c^{-1} \overline{d}^T \hat{S}_{22,1}^{-1} \overline{d}
\]

\[= c^{-1}(\bar{z}_1 - \bar{z}_2)^2 + c^{-1} \overline{d}^T S^{-1} \overline{d}\]
Decomposition of M.d. (contd.)

**Theorem.** If $\mu_1 = \mu_2$ then

$$\frac{D_{p+1}^2 - D_1^2}{(n-2) + D_1^2} = \frac{c^{-1} \mathbf{d}^T S^{-1} \mathbf{d}}{(n-2) + c^{-1} (z_1 - z_2)^2} \sim \frac{p}{n-2-p} F_{p,n-2-p}$$
Decomposition of M.d. (contd.)

Theorem. If $\mu_1 = \mu_2$ then

$$\frac{D_{p+1}^2 - D_1^2}{(n-2) + D_1^2} = \frac{c^{-1}d^T S^{-1}d}{(n-2) + c^{-1}(z_1 - z_2)^2} \sim \frac{p}{n-2-p} F_{p,n-2-p}$$

Corollary. Let $\mu_1 = \mu_2$ and let $\theta$ be the angle of separation for the normalized observations $\Sigma^{-1/2}x$. Then

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