STATISTICS 575
MULTIVARIATE ANALYSIS

Doug Wiens*

March 16, 2015

Contents

II INFERENCE ABOUT MEANS 32

Part I

INTRODUCTION

1. Introduction; sample statistics

- Intro to R – Table 1.3 in text. $n = 25$ lizards, $p = 3$ measurements on each: mass, snout-vent length (svl), hind limb span (hls). Do we need all three to describe this population, or will one or two, or a linear combination of the three, do just as well? R code is on course web site.

- $\mathbf{x}$ a $p \times 1$ random vector with elements X_i, from a population with mean $\mu_{p\times 1}$, covariance $\mathbf{\Sigma}_{p\times p}$:

$$\begin{aligned} E[\mathbf{x}] &= \mu, \text{ i.e. } E[X_i] = \mu_j \text{ for } j = 1, ..., p; \\ \text{cov}[\mathbf{x}] &= \mathbf{\Sigma} = \left(\sigma_{jk}\right), \text{ i.e.} \\ \text{cov}\left[X_j, X_k\right] &= \sigma_{jk} \text{ for } j, k = 1, .., p. \end{aligned}$$

Then $\text{corr}\left[X_j, X_k\right] = \sigma_{jk}/\sqrt{\sigma_{jj}\sigma_{kk}}$.

  - Useful identity:

$$\mathbf{\Sigma} = E\left[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\right] = E\left[\mathbf{x}\mathbf{x}^T\right] - \mu\mu^T.$$

- Gather a sample: $\mathbf{x}_1, .., \mathbf{x}_n$ independent observations from a population with mean $\mu_{p\times 1}$, covariance $\Sigma_{p\times p}$. So each is a realization of an $\mathbf{x}$ distributed in this manner. Make these the rows of a data matrix $\mathbf{X}$:

$$\mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}.$$

From this we can easily represent the <u>sample mean vector</u> and <u>sample covariance matrix</u>:
define $\mathbf{1}_n = (1, ..., 1)^T$: $n \times 1$, then

$$\begin{aligned} \bar{\mathbf{x}}_{p\times 1} &= \frac{1}{n}\sum_i \mathbf{x}_i = \frac{1}{n}\mathbf{X}^T\mathbf{1}_n, \\ \mathbf{S}_{p\times p} &= \frac{1}{n-1}\sum_i \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T. \end{aligned}$$

- Note

$$\begin{aligned}
(n-1)\mathbf{S} &= \left(\begin{array}{ccc} \mathbf{x}_1 - \bar{\mathbf{x}} & \cdots & \mathbf{x}_n - \bar{\mathbf{x}} \end{array} \right) \left(\begin{array}{c} (\mathbf{x}_1 - \bar{\mathbf{x}})^T \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})^T \end{array} \right) \\
&= (\cdots) \left(\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^T \right) \\
&= (\cdots) \left(\mathbf{X} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \mathbf{X} \right) \\
&= (\cdots) \left(\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \mathbf{X} \\
&= \left[\mathbf{X}^T \left(\mathbf{I}_n - \mathbf{J}_n \right) \right] \left[\left(\mathbf{I}_n - \mathbf{J}_n \right) \mathbf{X} \right],
\end{aligned}$$

where $\mathbf{J}_n = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ is symmetric and idempotent, hence so is $\mathbf{I}_n - \mathbf{J}_n$. Thus

$$(n-1)\mathbf{S} = \mathbf{X}^T \left(\mathbf{I}_n - \mathbf{J}_n \right) \mathbf{X},$$

and this can be continued as

$$\begin{aligned}
(n-1)\mathbf{S} &= \mathbf{X}^T \mathbf{X} - n \bar{\mathbf{x}} \bar{\mathbf{x}}^T \\
&= \sum_i \mathbf{x}_i \mathbf{x}_i^T - n \bar{\mathbf{x}} \bar{\mathbf{x}}^T. \qquad (1.1)
\end{aligned}$$

- These estimates are unbiased. Clearly $E\left[\bar{\mathbf{x}}\right]=\mu$; also cov$[\bar{\mathbf{x}}]=n^{-1}\mathbf{\Sigma}$:

$$\begin{aligned}
\text{cov}\left[\bar{\mathbf{x}}\right] &= E\left[\left(\bar{\mathbf{x}}-\mu\right)\left(\bar{\mathbf{x}}-\mu\right)^T\right] \\
&= E\left[\frac{1}{n}\sum_i\left(\mathbf{x}_i-\mu\right)\cdot\frac{1}{n}\sum_j\left(\mathbf{x}_j-\mu\right)^T\right] \\
&= \frac{1}{n^2}\sum_{i,j}E\left[\left(\mathbf{x}_i-\mu\right)\left(\mathbf{x}_j-\mu\right)^T\right] \\
\text{why?} &= \frac{1}{n^2}\sum_i E\left[\left(\mathbf{x}_i-\mu\right)\left(\mathbf{x}_i-\mu\right)^T\right] \\
&= \frac{1}{n}\mathbf{\Sigma}.
\end{aligned}$$

Then, by (1.1), we get

$$\begin{aligned}
E\left[(n-1)\mathbf{S}\right] &= \sum_i E\left[\mathbf{x}_i\mathbf{x}_i^T\right]-nE\left[\bar{\mathbf{x}}\bar{\mathbf{x}}^T\right] \\
&= n\left[\mathbf{\Sigma}+\mu\mu^T\right]-n\left[\frac{1}{n}\mathbf{\Sigma}+\mu\mu^T\right] \\
&= (n-1)\mathbf{\Sigma},
\end{aligned}$$

so $E\left[\mathbf{S}\right]=\mathbf{\Sigma}$.

- **Sample correlation matrix**. The population correlation coefficients $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ are the elements of the population correlation matrix

  $$\mathbf{P} = \mathbf{D}_{\sigma}^{-1/2}\mathbf{\Sigma}\mathbf{D}_{\sigma}^{-1/2},$$

  where $\mathbf{D}_{\sigma} = diag\,(\sigma_{11}, ..., \sigma_{pp})$. They are estimated by the corresponding elements of

  $$\mathbf{R} = \mathbf{D}_{s}^{-1/2}\mathbf{S}\mathbf{D}_{s}^{-1/2},$$

  where $\mathbf{D}_{s} = diag\,(S_{11}, ..., S_{pp})$, the diagonal matrix of sample variances.

2. Multivariate normality

- **Characteristic functions**: $\psi_{\mathbf{x}}(\mathbf{t}) = E\left[\exp\left(i\mathbf{t}^T\mathbf{x}\right)\right]$, ($\mathbf{t}$ a vector of real elements). Similar to moment generating functions, but the c.f. always exists.

  – Uniqueness: If $\psi_{\mathbf{x}}(\mathbf{t}) = \psi_{\mathbf{y}}(\mathbf{t})$ for all $\mathbf{t}$ then $\mathbf{x} \sim \mathbf{y}$.

  – Convergence: $\mathbf{x}^{(n)} \xrightarrow{L} \mathbf{x} \Leftrightarrow \psi_{\mathbf{x}^{(n)}}(\mathbf{t}) \to \psi_{\mathbf{x}}(\mathbf{t})$ for all $\mathbf{t}$. ($\xrightarrow{L}$ means ...)

- **Cramér-Wold device**:

$$\begin{aligned}
& \mathbf{x}^{(n)} \xrightarrow{L} \mathbf{x} \\
\Leftrightarrow\ & E\left[\exp\left(i\mathbf{t}^T\mathbf{x}^{(n)}\right)\right] \to E\left[\exp\left(i\mathbf{t}^T\mathbf{x}\right)\right] \text{ for all } \mathbf{t} \\
\Leftrightarrow\ & E\left[\exp\left(is\mathbf{t}^T\mathbf{x}^{(n)}\right)\right] \to E\left[\exp\left(is\mathbf{t}^T\mathbf{x}\right)\right] \text{ for all } s, \mathbf{t} \\
\Leftrightarrow\ & \mathbf{t}^T\mathbf{x}^{(n)} \xrightarrow{L} \mathbf{t}^T\mathbf{x} \text{ for all } \mathbf{t};
\end{aligned}$$

  i.e. we have convergence in law of $\mathbf{x}^{(n)}$ to $\mathbf{x}$ iff all linear combinations of $\mathbf{x}^{(n)}$ converge in law to those of $\mathbf{x}$.

- **Multivariate normality**. We adopt a roundabout definition, to handle the case in which the density might not exist due to a singular covariance matrix. First, we say that a univariate r.v. X has the $N\left(\mu,\sigma^2\right)$ distribution if the c.f. is

$$E\left[e^{itX}\right] = \exp\left\{it\mu - \frac{\sigma^2 t^2}{2}\right\}.$$

Then

$$E\left[e^{itX}\right] \text{ is the c.f. of a r.v. with}$$
$$\begin{cases} P\left(X=\mu\right)=1, & \text{if } \sigma^2 = 0, \\ \text{pdf } \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, & \text{if } \sigma^2 > 0, \end{cases}$$

so that these are the distributions (why?). If $\sigma^2 = 0$ then the "density" is concentrated at a single point μ ("Dirac's delta").

- Now let μ be a $p\times 1$ vector and $\mathbf{\Sigma}$ a $p\times p$ positive semidefinite matrix (i.e. $\mathbf{x}^T\mathbf{\Sigma}\mathbf{x} \geq 0$ for all $\mathbf{x}$). We write $\mathbf{\Sigma} \geq \mathbf{0}$. If $\mathbf{\Sigma}$ is positive definite ($\mathbf{\Sigma} > \mathbf{0}$), i.e. $\mathbf{x}^T\mathbf{\Sigma}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{\Sigma}$ is invertible.

- **Definition**: *We say that a r.vec.* $\mathbf{x}$ *has the multivariate normal* $N_p(\mu, \mathbf{\Sigma})$ *distribution if the c.f. is*

$$E\left[\exp\left(i\mathbf{t}^T\mathbf{x}\right)\right] = \exp\left\{i\mathbf{t}^T\mu - \frac{\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}}{2}\right\}.$$

- Putting all but one component of $\mathbf{t}$ equal to 0 yields the consequence that then $X_j \sim N\left(\mu_j, \sigma_j^2\right)$, where $\sigma_j^2 = \mathbf{\Sigma}_{jj}$. Comparing

$$\begin{aligned} & \frac{\partial}{\partial t_j} E\left[\exp\left(i\mathbf{t}^T\mathbf{x}\right)\right]_{|\mathbf{t}=\mathbf{0}} \\ = \ & E\left[iX_j \exp\left(i\mathbf{t}^T\mathbf{x}\right)\right]_{|\mathbf{t}=\mathbf{0}} = iE[X_j] \end{aligned}$$

to

$$\frac{\partial}{\partial t_j} \exp\left\{i\mathbf{t}^T\mu - \frac{\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}}{2}\right\}_{|\mathbf{t}=\mathbf{0}} = i\mu_j$$

yields $E[X_j] = \mu_j$ and similarly $COV\left[X_j, X_l\right] = \sigma_{jl}$. Thus $\mathbf{\Sigma}$ is the covariance matrix.

- If $\boldsymbol{\Sigma} > \mathbf{0}$ then there is a density

$$\begin{aligned} & \phi(\mathbf{x}; \mu, \boldsymbol{\Sigma}) \\ = \; & (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{(\mathbf{x}-\mu)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mu)}{2} \right\}. \end{aligned}$$

(What must be shown, to prove this?)

- If $\mathbf{x} \sim N_p(\mu_{\mathbf{x}}, \boldsymbol{\Sigma})$ then an arbitrary linear combination $\mathbf{c}^T\mathbf{x}$ has c.f.

$$\begin{aligned} & E\left[e^{it\mathbf{c}^T\mathbf{x}}\right] \\ = \; & E\left[\exp\left(i\mathbf{s}^T\mathbf{x}\right)\right]_{|\mathbf{s}=t\mathbf{c}} \\ = \; & \exp\left\{ i\mathbf{s}^T\mu_{\mathbf{x}} - \frac{\mathbf{s}^T\boldsymbol{\Sigma}\mathbf{s}}{2} \right\}_{|\mathbf{s}=t\mathbf{c}} \\ = \; & \exp\left\{ it\mu - \frac{\sigma^2 t^2}{2} \right\} \end{aligned}$$

with $\mu = \mathbf{c}^T\mu_{\mathbf{x}}$ and $\sigma^2 = \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$; thus

$$\mathbf{c}^T\mathbf{x} \sim N\left(\mathbf{c}^T\mu_{\mathbf{x}}, \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}\right).$$

(If $\boldsymbol{\Sigma}$ is singular then at least one of these univariate variances is zero.)

- Conversely, suppose that *every* linear combination is normally distributed. Then $\mathbf{x}$ is multivariate normal.

  Proof: Since $Y = \mathbf{t}^T\mathbf{x}$ is normal, it must be $N\left(\mathbf{t}^T\mu_{\mathbf{x}}, \mathbf{t}^T\mathbf{\Sigma}_{\mathbf{X}}\mathbf{t}\right)$, so that

$$\begin{aligned} E\left[\text{exp}\left(i\mathbf{t}^T\mathbf{x}\right)\right] &= E\left[e^{iY}\right] = \text{exp}\left\{i\mu_Y - \frac{\sigma_Y^2}{2}\right\} \\ &= \text{exp}\left\{i\mathbf{t}^T\mu_{\mathbf{x}} - \frac{\mathbf{t}^T\mathbf{\Sigma}_{\mathbf{X}}\mathbf{t}}{2}\right\}. \end{aligned}$$

  Thus $\mathbf{x}$ *is multivariate normal iff every linear combination is univariate normal*. This is the single most important property of this distribution.

- **Multivariate Central Limit Theorem**: Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be $p \times 1$ and i.i.d. with mean μ and covariance matrix $\mathbf{\Sigma}$. Then $\sqrt{n}\,(\bar{\mathbf{x}} - \mu) \xrightarrow{L} N_p\,(\mathbf{0}, \mathbf{\Sigma})$.

  Proof: We must show (why?) that

  $$\sqrt{n}\left(\mathbf{t}^T\bar{\mathbf{x}} - \mathbf{t}^T\mu\right) \xrightarrow{L} N\left(0, \mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\right)$$

  for every $\mathbf{t}$. This is the univariate CLT: put $Y_i = \mathbf{t}^T\mathbf{x}_i$; these are i.i.d. with mean $\mathbf{t}^T\mu$ and variance $\mathbf{t}^T\mathbf{\Sigma}\mathbf{t}$ and so $\sqrt{n}\left(\bar{Y} - \mathbf{t}^T\mu\right) \xrightarrow{L} N\left(0, \mathbf{t}^T\mathbf{\Sigma}\mathbf{t}\right)$. But $\bar{Y} = \mathbf{t}^T\bar{\mathbf{x}}$.

  - By this, many of the inferences discussed in the sequel, assuming normality, are at least asymptotically valid for non-normal populations.

- In the above if $\mathbf{\Sigma} > \mathbf{0}$ then (since convergence in law is preserved by continuous transformations) the continuous function
$$\left[\sqrt{n}\left(\bar{\mathbf{x}}-\mu\right)\right]^T \mathbf{\Sigma}^{-1}\left[\sqrt{n}\left(\bar{\mathbf{x}}-\mu\right)\right] \xrightarrow{L} \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$
where $\mathbf{y} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$. But $\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} = \mathbf{z}^T\mathbf{z}$ for $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{y} \sim N_p(\mathbf{0}, \mathbf{I})$. The elements Z_j are i.i.d. $N(0,1)$ and so
$$n\left(\bar{\mathbf{x}}-\mu\right)^T \mathbf{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\mu\right) \xrightarrow{L} \sum_{j=1}^{p} Z_j^2 \sim \chi_p^2.$$

- **Multivariate delta method**: If $\mathbf{f} : \mathbb{R}^p \to \mathbb{R}^q$ has continuous partial derivatives in a neighbourhood of μ, then
$$\sqrt{n}\left(\mathbf{f}\left(\bar{\mathbf{x}}\right)-\mathbf{f}\left(\mu\right)\right) \xrightarrow{L} N_q\left(\mathbf{0}, \mathbf{J}\mathbf{\Sigma}\mathbf{J}^T\right),$$
where
$$\mathbf{J}_{q\times p} = \frac{\partial \mathbf{f}}{\partial \mu} = \left(\frac{\partial f_i}{\partial x_j}_{|\mathbf{x}=\mu}\right)_{i,j}.$$

3. Marginal and conditional distributions; sufficient statistics

- Suppose that $\mathbf{x} \sim N_p(\mu, \mathbf{\Sigma})$. Using the c.f. approach, one shows (assigned) that an affine transformation $\mathbf{Ax} + \mathbf{b} \sim N_q\left(\mathbf{A}\mu + \mathbf{b}, \mathbf{A\Sigma A}^T\right)$ if $\mathbf{A}_{q\times p}$ and $\mathbf{b}_{q\times 1}$ are a matrix and vector with constant elements.

- Partition:

$$\begin{aligned}
\mathbf{x} &= \left(\begin{array}{c} \mathbf{x}_{(1)} : p_1 \times 1 \\ \mathbf{x}_{(2)} : p_2 \times 1 \end{array}\right),\ \mu = \left(\begin{array}{c} \mu_{(1)} \\ \mu_{(2)} \end{array}\right), \\
\mathbf{\Sigma} &= \left(\begin{array}{cc} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{array}\right);
\end{aligned}$$

note

$$\begin{aligned}
\mathbf{\Sigma}_{21} &= \mathbf{\Sigma}_{12}^T : p_2 \times p_1 = \text{cov}\left[\mathbf{x}_{(2)}, \mathbf{x}_{(1)}\right] \\
&= E\left[\left(\mathbf{x}_{(2)} - \mu_{(2)}\right)\left(\mathbf{x}_{(1)} - \mu_{(1)}\right)^T\right].
\end{aligned}$$

Then with $\mathbf{A} = \left(\begin{array}{cc} \mathbf{I}_{p_1} & \mathbf{0} \end{array}\right)$ or $\left(\begin{array}{cc} \mathbf{0} & \mathbf{I}_{p_2} \end{array}\right)$ in the above:

$$\mathbf{x}_{(j)} \sim N_{p_j}\left(\mu_{(j)}, \mathbf{\Sigma}_{jj}\right).$$

- A property of c.f.s is that $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are independent (iff their joint density factors into the product of marginal densities) iff their joint c.f. $E\left[\exp\left\{i\left(\mathbf{t}_{(1)}^T\mathbf{x}_{(1)}+\mathbf{t}_{(2)}^T\mathbf{x}_{(2)}\right)\right\}\right]$ factors into a function of $\mathbf{t}_{(1)}$ times a function of $\mathbf{t}_{(2)}$. A consequence (assigned) is that for *jointly normally distributed vectors*,

$$\mathbf{x}_{(1)} \text{ and } \mathbf{x}_{(2)} \text{ are independent } \Leftrightarrow \mathbf{\Sigma}_{12}=\mathbf{0}, \tag{3.1}$$

i.e. iff they are uncorrelated.

- Some useful matrix identities: For $\mathbf{\Sigma}>\mathbf{0}$, we have (derivations in text: exercises 4.10-4.12)

$$\mathbf{\Sigma}^{-1}=\begin{pmatrix}\mathbf{\Sigma}_{11\cdot2}^{-1} & -\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22\cdot1}^{-1}\\ * & \mathbf{\Sigma}_{22\cdot1}^{-1}\end{pmatrix},$$

where

$$\begin{aligned}\mathbf{\Sigma}_{11\cdot2} &= \mathbf{\Sigma}_{11}-\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21},\\ \mathbf{\Sigma}_{22\cdot1} &= \mathbf{\Sigma}_{22}-\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12},\end{aligned}$$

and $*$ denotes a block obtained by symmetry. Also

$$|\mathbf{\Sigma}|=|\mathbf{\Sigma}_{11}|\,|\mathbf{\Sigma}_{22\cdot1}|=|\mathbf{\Sigma}_{22}|\,|\mathbf{\Sigma}_{11\cdot2}|\,.$$

- **Conditional distributions**. Often we wish to make inferences about a subset of the population for which certain values are known, say $\mathbf{x}_{(2)} = \mathbf{c}_{(2)}$. This knowledge alters the distribution of $\mathbf{x}_{(1)}$. Given $\mathbf{x}_{(2)} = \mathbf{c}_{(2)}$, the distribution of $\mathbf{x}_{(1)}$ is Normal, with 'conditional mean'

  $$E\left[\mathbf{x}_{(1)}|\mathbf{x}_{(2)} = \mathbf{c}_{(2)}\right] = \mu_{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\left(\mathbf{c}_{(2)} - \mu_{(2)}\right)$$

  and 'conditional covariance'

  $$\text{cov}\left[\mathbf{x}_{(1)}|\mathbf{x}_{(2)} = \mathbf{c}_{(2)}\right] = \mathbf{\Sigma}_{11\cdot 2}.$$

  We write

  $$\mathbf{x}_{(1)}|\mathbf{x}_{(2)} \sim N_{p_1}\left(\mu_{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\left(\mathbf{x}_{(2)} - \mu_{(2)}\right), \mathbf{\Sigma}_{11\cdot 2}\right). \tag{3.2}$$

  Derivation: Define a r.vec. $\mathbf{y}$ by

  $$\begin{aligned} \begin{pmatrix} \mathbf{y}_{(1)} \\ \mathbf{y}_{(2)} \end{pmatrix} &= \begin{pmatrix} \mathbf{I}_{p_1} & -\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{p_2} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{(1)} \\ \mathbf{x}_{(2)} \end{pmatrix} \\ &\sim N_p\left(\begin{pmatrix} \mu_{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mu_{(2)} \\ \mu_{(2)} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{11\cdot 2} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{22} \end{pmatrix}\right). \end{aligned}$$

  Thus $\mathbf{y}_{(1)}$ is <u>independent</u> of $\mathbf{y}_{(2)} = \mathbf{x}_{(2)}$, so that $\mathbf{y}_{(1)}|\mathbf{x}_{(2)} \sim N_{p_1}\left(\mu_{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mu_{(2)}, \mathbf{\Sigma}_{11\cdot 2}\right)$ and so

$$\begin{aligned}\mathbf{x}_{(1)}|\mathbf{x}_{(2)} &= \left(\mathbf{y}_{(1)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_{(2)}\right)|\mathbf{x}_{(2)} \\ \sim\ & N_{p_1}\left(\mu_{(1)} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mu_{(2)} + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_{(2)}, \mathbf{\Sigma}_{11\cdot 2}\right),\end{aligned}$$

as required.

- Note that $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are independent $\Leftrightarrow$ the conditional distribution of $\mathbf{x}_{(1)}|\mathbf{x}_{(2)}$ does not depend on $\mathbf{x}_{(2)}$, and that, from (3.2), this happens $\Leftrightarrow \mathbf{\Sigma}_{12} = \mathbf{0}$, as we would expect from (3.1).

- We have $\mathbf{\Sigma}_{11} \geq \mathbf{\Sigma}_{11\cdot 2}$, in that the difference is p.s.d. – how? interpretation?

- Suppose $p_1 = 1$:

$$\begin{aligned}E\left[X_1|\mathbf{x}_{(2)}\right] &= \mu_{(1)} + \sigma_{12}^T\mathbf{\Sigma}_{22}^{-1}\left(\mathbf{x}_{(2)} - \mu_{(2)}\right) \\ &= \left(\mu_{(1)} - \sigma_{12}^T\mathbf{\Sigma}_{22}^{-1}\mu_{(2)}\right) + \sigma_{12}^T\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_{(2)} \\ &= \beta_0 + \sum_{i=1}^{p_2}\beta_i X_{2i};\end{aligned}$$

we say that ‘the regression of X_1 on $\mathbf{x}_{(2)}$ is linear’. Replacing $\mathbf{\Sigma}$ by $\mathbf{S}$, in order to estimate $E\left[X_1|\mathbf{x}_{(2)}\right]$, is equivalent to merely regressing the sampled values of X_1 on those of $\mathbf{x}_{(2)}$.

- **Conditional correlations**. As an example, suppose that $\mathbf{x}$ is trivariate normal, with covariance

$$\mathbf{\Sigma} = \left(\begin{array}{cc} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} & \begin{pmatrix} \sigma_{13} \\ \sigma_{23} \end{pmatrix} \\ \begin{pmatrix} \sigma_{31} & \sigma_{32} \end{pmatrix} & \sigma_{33} \end{array} \right).$$

Then the cov of (X_1, X_2) given X_3 is

$$\begin{aligned} \mathbf{\Sigma}_{11\cdot 2} &= \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} - \begin{pmatrix} \sigma_{13} \\ \sigma_{23} \end{pmatrix} \sigma_{33}^{-1} \begin{pmatrix} \sigma_{31} & \sigma_{32} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} - \frac{\sigma_{13}^2}{\sigma_{33}} & \sigma_{12} - \frac{\sigma_{13}\sigma_{32}}{\sigma_{33}} \\ * & \sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11}\left(1 - \rho_{13}^2\right) & \sigma_{12} - \frac{\sigma_{13}\sigma_{32}}{\sigma_{33}} \\ * & \sigma_{22}\left(1 - \rho_{23}^2\right) \end{pmatrix} \end{aligned}$$

so that the conditional correlation is

$$\begin{aligned} \rho_{12\cdot 3} &= \frac{\sigma_{12} - \frac{\sigma_{13}\sigma_{32}}{\sigma_{33}}}{\sqrt{\sigma_{11}\left(1-\rho_{13}^2\right)}\sqrt{\sigma_{22}\left(1-\rho_{23}^2\right)}} \\ &= \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}. \end{aligned}$$

These are estimated by replacing the population covariances by the sample covariances. There are startling examples in which $\rho_{12\cdot 3}$ and ρ_{12} have different signs ('Simpson's paradox' – see the Wikipedia article on the course web site for a discussion).

- Suppose that $\mathbf{x}_1, .., \mathbf{x}_n$ are independent observations from a $N_p(\mu, \mathbf{\Sigma})$ population with $\mathbf{\Sigma} > \mathbf{0}$. Then each has pdf

$$\begin{aligned} & \phi(\mathbf{x}; \mu, \mathbf{\Sigma}) \\ = \; & (2\pi)^{-p/2} |\mathbf{\Sigma}|^{-1/2} \exp\left\{ -\frac{(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)}{2} \right\}. \end{aligned}$$

The pdf of the sample is then

$$
\begin{aligned}
& \prod_{i=1}^{n} \phi(\mathbf{x}_i; \mu, \mathbf{\Sigma}) \\
= \;& (2\pi)^{-np/2} |\mathbf{\Sigma}|^{-n/2} \cdot \\
& \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu) \right\}.
\end{aligned}
$$

The sum is (with $\mathbf{S}_n = \frac{n-1}{n}\mathbf{S}$)

$$
\begin{aligned}
& \sum_{i=1}^{n} tr\left\{ (\mathbf{x}_i - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu) \right\} \\
= \;& \sum_{i=1}^{n} tr\left\{ \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu) (\mathbf{x}_i - \mu)^T \right\} \\
= \;& tr\left\{ \mathbf{\Sigma}^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \mu) (\mathbf{x}_i - \mu)^T \right\} \\
= \;& tr\left\{ \mathbf{\Sigma}^{-1} \left[n\mathbf{S}_n + n (\bar{\mathbf{x}} - \mu) (\bar{\mathbf{x}} - \mu)^T \right] \right\} \\
= \;& n \left[tr\mathbf{\Sigma}^{-1}\mathbf{S}_n + (\bar{\mathbf{x}} - \mu)^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \mu) \right].
\end{aligned}
\tag{3.3}
$$

- The resulting *likelihood function* is obtained by evaluating the pdf at the data, and is thus a function only of the unknown parameters:

$$\begin{aligned} L(\mu, \mathbf{\Sigma}) &= \prod_{i=1}^{n} \phi(\mathbf{x}_i; \mu, \mathbf{\Sigma}) \\ = & \ (2\pi)^{-np/2} |\mathbf{\Sigma}|^{-n/2} \cdot \\ & \exp\left\{-\frac{n}{2}\left[tr\mathbf{\Sigma}^{-1}\mathbf{S}_n + (\bar{\mathbf{x}} - \mu)^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \mu)\right]\right\}. \end{aligned}$$

  Thus $\bar{\mathbf{x}}$ and $\mathbf{S}_n$ are <u>sufficient statistics</u> – the likelihood depends on the data only through them. A consequence is that all inferences should be based only on these statistics.

- With probability one, $\mathbf{S}_n > \mathbf{0}$ if $\mathbf{\Sigma} > \mathbf{0}$. Why? If not, some quadratic form $= 0$:

$$0 = \mathbf{c}^T \mathbf{S}_n \mathbf{c} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{c}^T (\mathbf{x}_i - \bar{\mathbf{x}})\right)^2.$$

  Thus the random variables $\left\{\mathbf{c}^T \mathbf{x}_i\right\}_{i=1}^n$ are <u>constant</u> – they all equal $\mathbf{c}^T \bar{\mathbf{x}}$. Since they come from a distribution with positive variance $\mathbf{c}^T \mathbf{\Sigma} \mathbf{c}$, this is an event with probability 0.

- The most common method of estimation is maximum likelihood – the MLEs are the maximizers of the (log-) likelihood.

  Result: The MLEs of $(\mu, \mathbf{\Sigma})$ are $(\bar{\mathbf{x}}, \mathbf{S}_n)$.

  Proof: We are to show that $\log L(\mu, \mathbf{\Sigma}) \leq \log L(\bar{\mathbf{x}}, \mathbf{S}_n)$. But

$$\begin{aligned} & \log L(\bar{\mathbf{x}}, \mathbf{S}_n) - \log L(\mu, \mathbf{\Sigma}) \\ = & \frac{n}{2} \left\{ \begin{array}{c} \log|\mathbf{\Sigma}| - \log|\mathbf{S}_n| + tr\mathbf{\Sigma}^{-1}\mathbf{S}_n - p \\ + (\bar{\mathbf{x}} - \mu)^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \mu) \end{array} \right\}. \end{aligned}$$

  Thus it suffices to show that

$$tr\mathbf{\Sigma}^{-1}\mathbf{S}_n - \log|\mathbf{\Sigma}^{-1}\mathbf{S}_n| \geq p.$$

  Both in the trace and in the determinant we can replace $\mathbf{\Sigma}^{-1}\mathbf{S}_n$ by $\mathbf{\Sigma}^{-1/2}\mathbf{S}_n\mathbf{\Sigma}^{-1/2}$ (why?); then the trace is the sum, and the determinant is the product, of the (necessarily positive) eigenvalues $\{\lambda_i\}_{i=1}^p$ of this matrix. We are left having to show that

$$\sum_{i=1}^n (\lambda_i - \log\lambda_i) \geq p,$$

  which follows since the (convex) function $f(\lambda) = \lambda - \log\lambda$ $(\lambda > 0)$ is minimized at $\lambda = 1$ with $f(1) = 1$.

4. MLEs and their distributions

- **These MLEs are independently distributed**. Let $\mathbf{x}_1, .., \mathbf{x}_n$ be a $N_p(\mu, \boldsymbol{\Sigma})$ sample, and let $\mathbf{X}$ be the $n \times p$ data matrix,with rows $\mathbf{x}_i^T$. Let $\mathbf{Q}$ be an $n \times n$ orthogonal matrix with first row $\mathbf{q}_1^T = \frac{1}{\sqrt{n}}\mathbf{1}_n^T$ (how?); define $\mathbf{Y} = \mathbf{QX}$. The np elements of $\mathbf{X}$ are jointly normally distributed, hence so are those of $\mathbf{Y}$. The transpose of the i^{th} row of $\mathbf{Y}$ is $\mathbf{y}_i = \sum_k q_{ik}\mathbf{x}_k$; it follows that the distribution of any pair $\binom{\mathbf{y}_i}{\mathbf{y}_j}$ is

$$\begin{aligned} & N_{2p}\left(\left(\begin{array}{c} \left(\mathbf{q}_i^T\mathbf{1}_n\right)\mu \\ \left(\mathbf{q}_j^T\mathbf{1}_n\right)\mu \end{array}\right), \left(\begin{array}{cc} \left(\mathbf{q}_i^T\mathbf{q}_i\right)\boldsymbol{\Sigma} & \left(\mathbf{q}_i^T\mathbf{q}_j\right)\boldsymbol{\Sigma} \\ * & \left(\mathbf{q}_j^T\mathbf{q}_j\right)\boldsymbol{\Sigma} \end{array}\right)\right) \\ = \; & N_{2p}\left(\left(\begin{array}{c} \sqrt{n}\left(\mathbf{q}_i^T\mathbf{q}_1\right)\mu \\ \sqrt{n}\left(\mathbf{q}_j^T\mathbf{q}_1\right)\mu \end{array}\right), \left(\begin{array}{cc} \boldsymbol{\Sigma} & 0 \\ 0 & \boldsymbol{\Sigma} \end{array}\right)\right). \end{aligned}$$

Thus the $\{\mathbf{y}_i\}_{i=1}^n$ are independently and normally distributed. For $i = 1$ the mean is $\sqrt{n}\mu$ (and $\mathbf{y}_1 = \sqrt{n}\bar{\mathbf{x}}$); if $i \neq 1$ then $\mathbf{q}_i$ and $\mathbf{q}_1$ are orthogonal so the mean is $\mathbf{0}$. In summary:

$$\begin{aligned} \mathbf{y}_1 = \sqrt{n}\bar{\mathbf{x}} &\sim N_p\left(\sqrt{n}\mu, \boldsymbol{\Sigma}\right), \\ \mathbf{y}_2, ..., \mathbf{y}_n &\sim N_p\left(\mathbf{0}, \boldsymbol{\Sigma}\right). \end{aligned}$$

Note that $\mathbf{q}_1\mathbf{q}_1^T = \mathbf{J}_n$, so that if we partition $\mathbf{Q}$ as $\mathbf{Q} = \begin{pmatrix} \mathbf{q}_1^T \\ \mathbf{Q}_2^T \end{pmatrix}$, then $\mathbf{Q}_2\mathbf{Q}_2^T = \mathbf{I}_n - \mathbf{J}_n$ and

$$\begin{pmatrix} \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix} = \mathbf{Q}_2^T\mathbf{X}.$$

Now recall that

$$\begin{aligned} n\mathbf{S}_n &= \mathbf{X}^T(\mathbf{I}_n - \mathbf{J}_n)\mathbf{X} \\ &= \mathbf{X}^T\mathbf{Q}_2\mathbf{Q}_2^T\mathbf{X} \\ &= \sum_{i=2}^{n} \mathbf{y}_i\mathbf{y}_i^T. \end{aligned}$$

Thus:

- $\mathbf{S}_n$, which is a function only of $\mathbf{y}_2, ..., \mathbf{y}_n$, is independent of $\bar{\mathbf{x}} = \mathbf{y}_1/\sqrt{n}$.

- $\bar{\mathbf{x}} \sim N_p\left(\mu, \frac{1}{n}\mathbf{\Sigma}\right)$.

- $n\mathbf{S}_n$ ($= (n-1)\mathrm{S}$) is distributed as $\sum_{i=1}^{n-1} \mathbf{z}_i\mathbf{z}_i^T$, where the $\mathbf{z}_i$ are i.i.d. $N_p(\mathbf{0}, \mathbf{\Sigma})$ r.vecs.

- The matrix $\mathbf{W} = \sum_{i=1}^{m} \mathbf{z}_i \mathbf{z}_i^T$, where the $\mathbf{z}_i$ are i.i.d. $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ r.vecs, has the 'Wishart $W_p(m, \boldsymbol{\Sigma})$ distribution on m df and parameter $\boldsymbol{\Sigma}$'. (This means the distribution of the $p(p+1)/2$ distinct elements of $\mathbf{W}$ – pdf is given in text.) Some immediate consequences:

  - If $p = 1$ then $\mathbf{z}_i \mathbf{z}_i^T = \sigma^2 Z_i^2$, where Z_i is $N(0, 1)$, hence $W \sim \sigma^2 \chi_m^2$.

  - $\boldsymbol{\Sigma}^{-1/2} \mathbf{W} \boldsymbol{\Sigma}^{-1/2} \sim W_p(m, \mathbf{I})$.

  - If $\mathbf{W}_1 \sim W_p(m_1, \boldsymbol{\Sigma})$, independently of $\mathbf{W}_2 \sim W_p(m_2, \boldsymbol{\Sigma})$, then $\mathbf{W}_1 + \mathbf{W}_2 \sim W_p(m_1 + m_2, \boldsymbol{\Sigma})$.

  - If $\mathbf{W} \sim W_p(m, \boldsymbol{\Sigma})$ then $E[\mathbf{W}] = m\boldsymbol{\Sigma}$.

- **Noncentral distributions**. If $Z_1, ..., Z_m$ are independent, with $Z_i \sim N(\mu_i, \sigma^2)$, then $X_i = Z_i/\sigma \sim N(\nu_i, 1)$ (where $\nu_i = \mu_i/\sigma$) and $X^2 = \sum_{i=1}^{m} X_i^2$ has the 'non-central χ_m^2' distribution, with 'non-centrality parameter' (ncp) $\lambda^2 = \sum_{i=1}^{m} \nu_i^2$.

We write $X^2 \sim \chi^2_m\left(\lambda^2\right)$. If $\lambda^2 = 0$ this is the well-known central χ^2_m distribution. If $X^2 \sim \chi^2_m\left(\lambda^2_1\right)$, independent of $Y^2 \sim \chi^2_n\left(\lambda^2_2\right)$ then

$$\frac{X^2/m}{Y^2/n} \sim F^m_n\left(\lambda^2_1, \lambda^2_2\right),$$

the 'doubly non-central F^m_n' distribution. If $\lambda^2_2 = 0$ it is 'singly non-central', and this is the case most commonly encountered. If as well $\lambda^2_1 = 0$ this is the usual F^m_n distribution.

- A combination of Assignment 1 problem 4(b), and its continuation on the course web site, shows that if $X^2 \sim \chi^2_m\left(\lambda^2\right)$ then we can represent it as a *central* chi-square with *random degrees of freedom*:

  $$X^2 \sim \chi^2_{2K+m},$$

  where K is a Poisson r.v. with mean $\lambda^2/2$ (so $P\left(K = 0\right) = 1$ if $\lambda = 0$).

Part II

INFERENCE ABOUT MEANS

5. Inferences on one mean vector

- Sample $\mathbf{x}_1, ..., \mathbf{x}_n \sim i.i.d.\ N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$. We might seek:
  - Confidence regions (or intervals) on $\boldsymbol{\mu}$,
  - Prediction regions for a new $\mathbf{x}_*$ from the same population.

- **Sufficient statistics**: $\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}\right)$, independently of $(n-1)\,\mathbf{S} \sim W_p(\boldsymbol{\Sigma}, n-1)$.

- Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. [Why? 1. $\boldsymbol{\mu}$ a vector of linear functions of another mean, e.g. in repeated measures? 2. Invert, to get a confidence region?].

- In multivariate analysis the most common test is the likelihood ratio test. Here we look at the ratio of (maximized) likelihood functions, with and without assuming the truth of the hypothesis:

$$\Lambda = \frac{\max_{H_0} L(\mu, \mathbf{\Sigma})}{\max_{H_0 \cup H_1} L(\mu, \mathbf{\Sigma})}.$$

If H_0 holds, we expect Λ to be near 1. Small values of Λ arise if the maximum of the likelihood is attained at values not in agreement with H_0; if Λ is too small we reject.

In the current application

$$\begin{aligned} &L(\mu, \mathbf{\Sigma}) = (2\pi)^{-np/2} |\mathbf{\Sigma}|^{-n/2} \cdot \\ &\exp\left\{-\frac{n}{2}\left[tr\mathbf{\Sigma}^{-1}\mathbf{S}_n + (\bar{\mathbf{x}} - \mu)^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \mu)\right]\right\}. \end{aligned}$$

The unconditional maximizers are $\hat{\mu} = \bar{\mathbf{x}}$ and $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$, so that

$$\max_{H_0 \cup H_1} L(\mu, \mathbf{\Sigma}) = L(\bar{\mathbf{x}}, \mathbf{S}_n) = (2\pi e)^{-np/2} |\mathbf{S}_n|^{-n/2}.$$

Under H_0, the likelihood is

$$\begin{aligned} &L(\mu_0, \mathbf{\Sigma}) = (2\pi)^{-np/2} |\mathbf{\Sigma}|^{-n/2} \\ &\cdot \exp\left\{-\frac{n}{2}\left[tr\mathbf{\Sigma}^{-1}\mathbf{S}_n + (\bar{\mathbf{x}} - \mu_0)^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{x}} - \mu_0)\right]\right\} \\ = \;\; &(2\pi)^{-np/2} |\mathbf{\Sigma}|^{-n/2} \exp\left\{-\frac{n}{2}\left[tr\mathbf{\Sigma}^{-1}\tilde{S}_n\right]\right\}, \end{aligned}$$

where

$$\tilde{S}_n = \mathbf{S}_n + (\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)^T .$$

Exactly as the unrestricted MLE $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$ was obtained, we get the restricted MLE $\hat{\mathbf{\Sigma}}_0 = \tilde{S}_n$, with

$$\max_{\mathbf{\Sigma}} L(\mu_0, \mathbf{\Sigma}) = L\left(\mu_0, \tilde{S}_n\right) = (2\pi e)^{-np/2} |\tilde{S}_n|^{-n/2}.$$

Thus

$$\Lambda = \frac{L\left(\mu_0, \hat{\mathbf{\Sigma}}_0\right)}{L\left(\bar{\mathbf{x}}, \hat{\mathbf{\Sigma}}\right)} = \frac{(2\pi e)^{-np/2} |\tilde{S}_n|^{-n/2}}{(2\pi e)^{-np/2} |\mathbf{S}_n|^{-n/2}} = \left(\frac{|\tilde{S}_n|}{|\mathbf{S}_n|}\right)^{-n/2} .$$

Note that

$$\begin{aligned} |\tilde{S}_n| &= |\mathbf{S}_n + (\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)^T| \\ &= |\mathbf{S}_n||\mathbf{I}_p + \mathbf{S}_n^{-1}(\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)^T| \\ &= |\mathbf{S}_n|\left(1 + (\bar{\mathbf{x}} - \mu_0)^T \mathbf{S}_n^{-1}(\bar{\mathbf{x}} - \mu_0)\right). \end{aligned}$$

How? Calculate

$$\left|\begin{pmatrix} \mathbf{I}_p & \mathbf{a} \\ \mathbf{b}^T & 1 \end{pmatrix}\right|$$

in two ways, to get $\left|\mathbf{I}_p - \mathbf{a}\mathbf{b}^T\right| = 1 - \mathbf{b}^T\mathbf{a}$.

We have now obtained

$$\Lambda = \left(1 + (\bar{\mathbf{x}} - \mu_0)^T \mathbf{S}_n^{-1}(\bar{\mathbf{x}} - \mu_0)\right)^{-n/2} .$$

In terms of

$$T^2 = n\left(\bar{\mathbf{x}} - \mu_0\right)^T \mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \mu_0\right),$$

this is

$$\Lambda = \left(1 + \frac{T^2}{n-1}\right)^{-n/2}$$

so that rejecting for small Λ is equivalent to rejecting for large T^2. [A connection with $p = 1$: in the univariate case, $t^2 = \left(\sqrt{n}\left(\bar{x} - \mu_0\right)/s\right)^2$.]

- **Theorem**: If $\mathbf{z} \sim N_p\left(0, \mathbf{I}\right)$, independently of $\mathbf{V} \sim W_p\left(n-1, \mathbf{I}\right)$, then $\mathbf{z}^T\left(\frac{\mathbf{V}}{n-1}\right)^{-1}\mathbf{z} \sim T^2_{p,n-p}$ – termed "Hotelling's T^2". The distribution is

  $$\frac{n-p}{p}\frac{T^2}{n-1} \sim F^p_{n-p}.$$

  If instead $\mathbf{z}$ has mean vector ν, the F is non-central, with ncp $\lambda^2 = \nu^T\nu$.

  - If $p = 1$ we get $t^2 \sim F^1_{n-1}$; obvious since in this case $t \sim t_{n-1}$.

Harold Hotelling (September 29, 1895 – December 26, 1973) was a mathematical statistician and an influential economic theorist. He was Associate Professor of Mathematics at Stanford University from 1927 until 1931, a member of the faculty of Columbia University from 1931 until 1946, and a Professor of Mathematical Statistics at the University of North Carolina at Chapel Hill from 1946 until his death. A street in Chapel Hill bears his name.

- To link this T^2-distribution to the T^2-statistic arising in the LR test, put

$$\begin{aligned}
\mathbf{z} &= \sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}}-\mu_0) \\
&\sim N_p\left(\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\mu-\mu_0),\mathbf{I}\right), \\
\mathbf{V} &= \boldsymbol{\Sigma}^{-1/2}\left[(n-1)\,\mathbf{S}\right]\boldsymbol{\Sigma}^{-1/2} \\
&\sim W_p(n-1,\mathbf{I}).
\end{aligned}$$

  Then

$$\begin{aligned}
\mathbf{z}^T\left(\frac{\mathbf{V}}{n-1}\right)^{-1}\mathbf{z} &= n(\bar{\mathbf{x}}-\mu_0)^T\mathbf{S}^{-1}(\bar{\mathbf{x}}-\mu_0)=T^2, \\
\lambda^2 &= n(\mu-\mu_0)^T\boldsymbol{\Sigma}^{-1}(\mu-\mu_0).
\end{aligned}$$

  More simply,

$$\begin{aligned}
T^2 &= (\bar{\mathbf{x}}-\mu_0)^T\left[\widehat{\mathsf{cov}(\bar{\mathbf{x}})}\right]^{-1}(\bar{\mathbf{x}}-\mu_0) \\
&\sim dfS\frac{df1}{df2}F^{df1}_{df2}\left(\lambda^2\right), \qquad (5.1)
\end{aligned}$$

  where

  - dfS is the df associated $(n-1)$ with the Wishart distribution of the (unbiased) covariance estimate,

- $df1$ is the dimension (p) of the cov estimate,
- $df1 + df2 = dfS + 1$,
- $\lambda^2 = T^2$ with all estimates replaced by their expectations.

- This test also arising from the *union-intersection principle* – another multivariate testing procedure. Here, we argue that H_0 holds iff $H_{\mathbf{a}} : \mathbf{a}^T\mu = \mathbf{a}^T\mu_0$ holds for all $\mathbf{a}$; hence we should reject iff any of the univariate hypotheses $H_{\mathbf{a}}$ are rejected. But since $\mathbf{a}^T\bar{\mathbf{x}} \sim N\left(\mathbf{a}^T\mu, n^{-1}\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}\right)$, and the sample variance of the $\mathbf{a}^T\mathbf{x}_i$ is $\mathbf{a}^T\mathbf{S}\mathbf{a}$, we reject $H_{\mathbf{a}}$ for large values of $\left|\frac{\sqrt{n}(\mathbf{a}^T\bar{\mathbf{x}}-\mathbf{a}^T\mu_0)}{\sqrt{\mathbf{a}^T\mathbf{S}\mathbf{a}}}\right| \sim |t_{n-1}|$. Thus H_0 is rejected iff

$$\max_{\mathbf{a}}\left|\frac{\sqrt{n}\left(\mathbf{a}^T\bar{\mathbf{x}} - \mathbf{a}^T\mu_0\right)}{\sqrt{\mathbf{a}^T\mathbf{S}\mathbf{a}}}\right| > c,$$

where c is chosen appropriately. But $\left|\frac{\sqrt{n}(\mathbf{a}^T\bar{\mathbf{x}}-\mathbf{a}^T\mu_0)}{\sqrt{\mathbf{a}^T\mathbf{S}\mathbf{a}}}\right|$ is maximized (assigned) by

$$\mathbf{a} = \mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \mu_0\right) \tag{5.2}$$

(the direction least in agreement with H_0), with maximum value $\sqrt{T^2}$.

- **Confidence regions**: invert a level α test to get a $100\,(1-\alpha)\,\%$ confidence region
$$\left\{\mu_0 | \frac{n-p}{p}\frac{T^2}{n-1} \leq F^p_{n-p}(\alpha)\right\}.$$
This is the ellipsoid
$$\left\{\begin{array}{c} \mu_0 | n\left(\bar{\mathbf{x}}-\mu_0\right)^T \mathbf{S}^{-1}\left(\bar{\mathbf{x}}-\mu_0\right) \\ \leq \frac{n-1}{n-p} p F^p_{n-p}(\alpha) \overset{def}{=} c^2 \end{array}\right\},$$
centred at $\bar{\mathbf{x}}$. The i^{th} semi-axis has length $c\sqrt{\lambda_i}$ in the direction $\mathbf{v}_i$, where $\{\lambda_i\}$ and $\{\mathbf{v}_i\}$ are the eigenvalues and corresponding eigenvectors of $\mathbf{S}/n$.

- A convenient representation: Write $\mathbf{S} = \mathbf{U}^T\mathbf{U}$, where $\mathbf{U}$ is upper triangular. Define
$$\mathbf{z} = -\frac{\sqrt{n}}{c}\mathbf{U}^{-T}\left(\bar{\mathbf{x}}-\mu_0\right).$$
Then the ellipsoid is $\left\{\mu_0 = \bar{\mathbf{x}} + \frac{c}{\sqrt{n}}\mathbf{U}^T\mathbf{z} \mid \|\mathbf{z}\| \leq 1\right\}$.

- Individual confidence interval on $\psi = \mathbf{a}^T\mu$:

  $$\begin{aligned} \hat{\psi} = \mathbf{a}^T\bar{\mathbf{x}} &\sim N(\psi, \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}/n), \\ \mathbf{a}^T\mathbf{S}\mathbf{a}/\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} &\sim \chi^2_{n-1}/(n-1), \end{aligned}$$

  (you should now be able to establish this last statement); thus

  $$\mathbf{a}^T(\bar{\mathbf{x}} - \mu) \bigg/ \sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}} = \frac{\hat{\psi} - \psi}{s_{\hat{\psi}}} \sim t_{n-1}.$$

  So a $100(1-\alpha)\%$ CI is

  $$\hat{\psi} \pm t_{n-1}(\alpha/2)\, s_{\hat{\psi}}.$$

  If we compute m of these, using $t_{n-1}\left(\frac{\alpha}{2m}\right)$ each time, then the overall level of confidence is at least $1-\alpha$ – ‘Bonferroni intervals’: before constructing them,

  $$\begin{aligned} & P\,(\text{all will be correct}) \\ = \ & 1 - P\left(\cup_{i=1}^m i^{th} \text{ will be wrong}\right) \\ \geq \ & \cdots = 1 - \alpha. \end{aligned}$$

- For 'data snooping' we want simultaneous CIs: we seek 'c' such that, before sampling,

$$\begin{aligned}1-\alpha &= P\left(\left|\frac{\mathbf{a}^T\left(\bar{\mathbf{x}}-\mu\right)}{\sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}}}\right| \leq c \text{ for all } \mathbf{a}\right) \\ &= P\left(\max_{\mathbf{a}}\left|\frac{\sqrt{n}\mathbf{a}^T\left(\bar{\mathbf{x}}-\mu\right)}{\sqrt{\mathbf{a}^T\mathbf{S}\mathbf{a}}}\right| \leq c\right) \\ &= P\left(T^2 \leq c^2\right),\end{aligned}$$

(the last equality uses (5.2)) so $c^2 = \frac{n-1}{n-p}pF_{n-p}^{p}(\alpha)$. The CIs are

$$\mathbf{a}^T\bar{\mathbf{x}} \pm c\sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}}.$$

6. Inferences on one mean vector cont'd

- College scores data set – Table 5.2; R code on course web site.
  - Check marginal normality – p-values obtained from the Shapiro-Wilk test, which compares the observed and expected sample quantiles, under the assumption of normality. R computes the p-values for this test; another test, which is almost as powerful, is based on the correlation between the observed and expected quantiles.
  - A possible way to check multivariate normality: if the $\mathbf{x}_i$ are $N_p(\mu, \mathbf{\Sigma})$ then
  $$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \overset{d}{\approx} \chi_p^2.$$
  Plot the ordered values $d_{(1)}^2 < \cdots < d_{(n)}^2$ against the quantiles $\chi_p^2\left(\frac{i-.5}{n}\right)$ (lower tail).

- Bonferroni intervals on the marginal means: $\hat{\psi} \pm t_{n-1}\left(\frac{\alpha}{2m}\right) s_{\hat{\psi}}$ with $m = 3$, $\hat{\psi} = \bar{\mathbf{x}}_i$, $s_{\hat{\psi}} = \sqrt{\mathbf{S}_{ii}/n}$.

- Simultaneous CIs:

$$\mathbf{a}^T\bar{\mathbf{x}} \pm c\sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}},$$

  with $c^2 = (n-1)\frac{p}{n-p}F^p_{n-p}(\alpha)$; the three means correspond to $\mathbf{a}^T = (1,0,0)\,,(0,1,0),(0,0,1)$. Compare with Bonferroni intervals.

- Plot bivariate confidence ellipse:

$$\left\{\mu = \bar{\mathbf{x}} + \frac{c}{\sqrt{n}}\mathbf{U}^T\mathbf{z} \mid \|\mathbf{z}\| \leq \mathbf{1}\right\},$$

  where $\mathbf{U}^T\mathbf{U} = \mathbf{S}$ (U `=` `chol`($\mathbf{S}$) on R).

Martin Wilk (Dec. 18, 1922 - Feb. 19, 2013). Former Chief Statistician of Canada; Officer of the Order of Canada. See the obituary on the course web site.

- **Prediction of new observations**: Suppose we monitor a process, making measurements $\mathbf{x}_1, ..., \mathbf{x}_n \sim i.i.d.\ N_p(\mu, \Sigma)$. These are known to be obtained when the process is 'in control'. A future observation $\mathbf{x}_{new}$ will be made; in order to see if it is in control we might construct a region within which it will fall with high probability if in control. Thus assume that $\mathbf{x}_{new} \sim N_p(\mu, \Sigma)$, then

$$\begin{aligned} \mathbf{x}_{new} - \bar{\mathbf{x}} &\sim N_p\left(0, \left(1 + \frac{1}{n}\right)\Sigma\right), \\ \text{ind. of } \mathbf{S} &\sim W_p(n-1, \Sigma) / (n-1). \end{aligned}$$

Hence

$$\begin{aligned} T^2 &= (\mathbf{x}_{new}-\bar{\mathbf{x}})^T \left\{\widehat{cov}\left[\mathbf{x}_{new}-\bar{\mathbf{x}}\right]\right\}^{-1} (\mathbf{x}_{new}-\bar{\mathbf{x}}) \\ &= (\mathbf{x}_{new}-\bar{\mathbf{x}})^T \left\{\left(1+\frac{1}{n}\right)\mathbf{S}\right\}^{-1} (\mathbf{x}_{new}-\bar{\mathbf{x}}) \\ &= \frac{n}{n+1}(\mathbf{x}_{new}-\bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_{new}-\bar{\mathbf{x}}) \\ &\sim dfS\frac{df1}{df2}F_{df2}^{df1} = (n-1)\frac{p}{n-p}F_{n-p}^{p} \end{aligned}$$

since $dfS = n-1$, $df1 = p$, $df1+df2 = dfS+1 = n$. Thus

$$1-\alpha = P\left(\begin{array}{c} (\mathbf{x}_{new}-\bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_{new}-\bar{\mathbf{x}}) \\ \leq \frac{n+1}{n}(n-1)\frac{p}{n-p}F_{n-p}^{p}(\alpha) \end{array} \right).$$

If this inequality fails the process may be declared out of control. Note that the computations are identical to those for a confidence ellipsoid, with c^2 replaced by $\frac{n+1}{n}c^2$. [R code for a bivariate example is on the course web site.]

- **Asymptotic inferences**. Suppose that the underlying sample is possibly non-normal, but the sample size is large. As $n \to \infty$ we have that $\mathbf{S} \xrightarrow{pr} \boldsymbol{\Sigma}$ and so, as $\sqrt{n}\,(\bar{\mathbf{x}} - \mu) \xrightarrow{L} \mathbf{z} \sim N_p\,(\mathbf{0}, \boldsymbol{\Sigma})$, we have

$$\begin{aligned} T^2 &= n\,(\bar{\mathbf{x}} - \mu)^T\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}} - \mu) \\ &= \left\| \mathbf{S}^{-1/2} \cdot \sqrt{n}\,(\bar{\mathbf{x}} - \mu) \right\|^2 \\ &\xrightarrow{L} \left\| \boldsymbol{\Sigma}^{-1/2} \mathbf{z} \right\|^2 \\ &\sim \chi^2_p. \end{aligned}$$

To apply this we need not change any of the methods described above, we merely use percentage points derived from the χ^2_p distribution, rather than those of $(n-1)\,\frac{p}{n-p} F^p_{n-p}$. Note also that

$$\begin{aligned} (n-1)\,\frac{p}{n-p} F^p_{n-p} &\sim (n-1)\,\frac{p}{n-p} \frac{\chi^2_p/p}{\chi^2_{n-p}/\,(n-p)} \\ &= \frac{n-1}{n-p} \frac{\chi^2_p}{\chi^2_{n-p}/\,(n-p)} \\ &\xrightarrow{L} \chi^2_p, \end{aligned}$$

since $\frac{n-1}{n-p} \to 1$ and, with $Z_1, ..., Z_{n-p} \overset{ind.}{\sim} N(0,1)$,

$$\frac{\chi^2_{n-p}}{n-p} \sim \frac{1}{n-p}\sum_{i=1}^{n-p} Z_i^2 \xrightarrow{pr} E\left[Z_1^2\right] = 1,$$

by the Weak Law of Large Numbers.

- There is a general large-sample approximation to the distribution of the LR statistic:

  $$-2\log\Lambda \xrightarrow{L} \chi^2_{df} \text{ as } n \to \infty,$$

  under H_0, where df is the reduction in the dimension of the parameter space, resulting from the null hypothesis (typically this equals the number of parameters being tested). There are numerous improvements to this, collectively referred to as *Barlett corrections*.

- See the code on the web site for a test of the hypothesis that the mean of the college scores population is $\mu_0 = (500, 50, 25)^T$, using both the exact method (under normality; $p = .00041$) and the large sample approximation ($p = .00032$).

7. Paired comparisons; repeated measures designs

- Some typical situations in which we might wish to make inferences about one or two population means:

  - Take n samples of water, split each into two, have one analyzed by one lab and the other by another lab (Example 6.1 in text). Each analysis results in $p = 2$ measurements (on 'BOD' – biochemical oxygen demand, and 'SS' – suspended solids); hence two matched samples $\{\mathbf{x}_{1i}\}$, $\{\mathbf{x}_{2i}\}$ and two $p \times 1$ mean vectors μ_1, μ_2. There might be two different covariance structures, but if the object of the study is to discern differences between the labs we are interested only in $\mu_d = \mu_1 - \mu_2$, estimated by the difference $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. So the relevant data are the differences $\mathbf{d}_i = \mathbf{x}_{1i} - \mathbf{x}_{2i}$, with mean μ_d estimated by the average of the $\mathbf{d}_i$, and some covariance estimated by the sample covariance of the $\mathbf{d}_i$. <u>One sample problem</u>.

– Make a number of measurements on each individual; the resulting outcomes will be dependent: ‘repeated measures’. Example 6.2 in text: administer $p = 4$ treatments to each of $n = 19$ dogs (asleep; all have been drugged). Purpose is to study the effect of halothane (H) and CO2 on heartbeat. Factorial structure: measure heartbeat after administering CO2 at a high or low level, and with and without treating with H.

$$\mu = \begin{pmatrix} \text{mean heartbeat, high CO2, no H} \\ \text{mean heartbeat, low CO2, no H} \\ \text{mean heartbeat, high CO2, with H} \\ \text{mean heartbeat, low CO2, with H} \end{pmatrix}.$$

Interesting ‘contrasts’:

$$\begin{aligned} (\mu_3 + \mu_4) - (\mu_1 + \mu_2) &= \text{overall effect of H}, \\ (\mu_1 + \mu_3) - (\mu_2 + \mu 4) &= \text{overall effect of CO2}, \\ (\mu_1 + \mu_4) - (\mu_2 + \mu_3) &= \text{interaction effect}; \end{aligned}$$

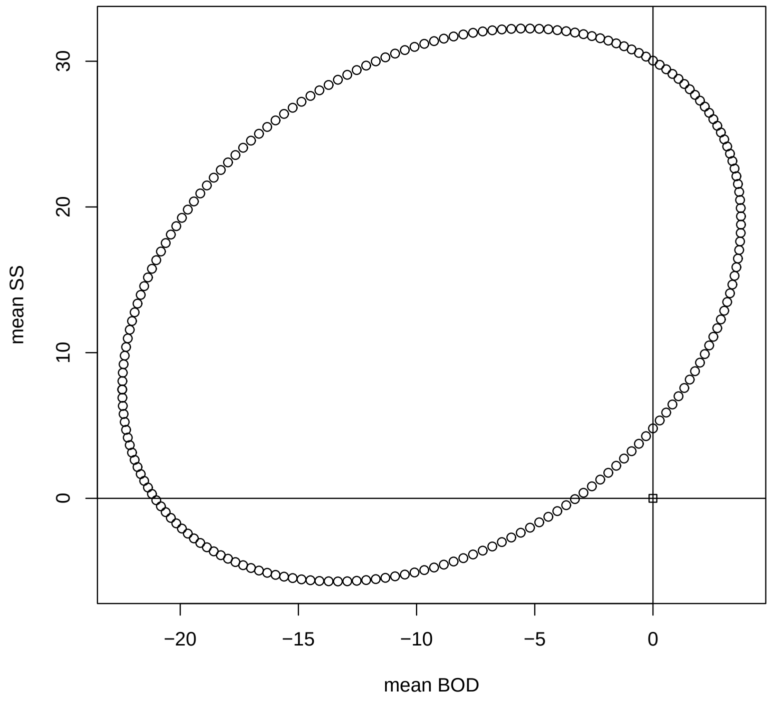
i.e. interest is on $\mathbf{C}\mu$, where

$$\mathbf{C} = \begin{pmatrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

is a ‘contrast matrix’ – rows are orthogonal to each other and to 1_4^T – they sum to 0. We want inferences about a single mean vector (that of $\mathbf{Cx}$).

- Two or more independent populations. Example 6.3 text: Fifty bars of soap are manufactured by one process, resulting in bivariate data $\{\mathbf{x}_{1i}\}$ (measures of lather and mildness) with mean μ_1, covariance $\mathbf{\Sigma}_1$. Fifty others are manufactured by another process, resulting in $\mathbf{x}_{2i} \sim (\mu_2, \mathbf{\Sigma}_2)$. It is reasonable to assume that the two samples are independent; we want to compare the two mean vectors. (In Example 6.1 above this independence of the two samples would not seem reasonable – why not? – but the problem is avoided by taking differences.)
  This is the subject of the next class.

- Example 6.1. See R code on course web site.



Confidence ellipsoid on mean between-lab differences in (BOD, SS); (0,0) shown.

– Note that $\mathbf{0} \notin$ the 95% confidence ellipse, so $H_0 : \mu_d = \mathbf{0}$ will be rejected at $\alpha = .05$. This means that at least one univariate $H_0 : \mathbf{a}^T \mu_d = 0$ will be rejected (why?). However, 0 is in each univariate interval (simultaneous or Bonferroni), meaning only that

$$H_0 : \mathbf{a}^T \mu_d = 0$$

is plausible both for $\mathbf{a} = (1,0)^T$ and $\mathbf{a} = (0,1)^T$.

```
 Bonferroni intervals on the two marginal means

          lower       dbar     upper
BOD -20.573107 -9.363636  1.845835
SS   -2.974903 13.272727 29.520358
```

In carrying out the experiment, in each of the n cases the <u>same</u> water was split into 2 samples, with each being sent to a different lab. This means that $\mathbf{x}_{1i}, \mathbf{x}_{2i}$ would be highly correlated. We expect that the components would be positively correlated, and so this matching reduces the variance:

$$\text{var}\left[X_1 - X_2\right] = \text{var}\left[X_1\right] + \text{var}\left[X_2\right] - 2\text{cov}\left[X_1, X_2\right].$$

- Example 6.2. If $\mathbf{x} \sim N_p(\mu, \mathbf{\Sigma})$ and $\mathbf{C}$ is a $p-1 \times p$ contrast matrix, then the statement that $\mathbf{C}\mu = \mathbf{0}$ is equivalent to the statement that all elements of μ are equal (to each other).

  - This is because $\mathbf{C}\mu = \mathbf{0}$ iff $\mu \perp$ every row of $\mathbf{C}$ (recall there are $p-1$ of these, and they are mutually orthogonal) iff μ lies in the orthogonal complement of the row space of $\mathbf{C}$, which is of dimension one with basis $\{1_p\}$.

  - More algebraically:
$$\mathbf{Q} = \begin{pmatrix} 1_p^T \\ \mathbf{C} \end{pmatrix}$$
satisfies $\mathbf{Q}^T\mathbf{Q} = p\,\mathbf{I}_p$, so that
$$\mathbf{I}_p = p^{-1}\mathbf{Q}^T\mathbf{Q} = \mathbf{J}_p + p^{-1}\mathbf{C}^T\mathbf{C};$$
hence
$$\begin{aligned} \mathbf{C}\mu = \mathbf{0} \ \Leftrightarrow p^{-1} \ & \mu^T\mathbf{C}^T\mathbf{C}\mu = 0 \\ \Leftrightarrow \quad & \mu^T(\mathbf{I}_p - \mathbf{J}_p)\mu = 0 \\ \Leftrightarrow \quad & \sum_{j=1}^{p}\left(\mu_j - \bar{\mu}\right)^2 = 0. \end{aligned}$$

– To make inferences on $\mathbf{C}\mu$ we can transform the data from $\{\mathbf{x}_i\}_{i=1}^n$ to $\{\mathbf{y}_i = \mathbf{C}\mathbf{x}_i\}_{i=1}^n \sim N_q\left(\mathbf{C}\mu, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T\right)$ with $q = p - 1$. Then

$$\begin{aligned} \bar{\mathbf{y}} &= \mathbf{C}\bar{\mathbf{x}}, \\ \mathbf{S}_{\mathbf{y}} &= \mathbf{C}\mathbf{S}_{\mathbf{x}}\mathbf{C}^T \end{aligned}$$

and so

$$\begin{aligned} T^2 &= n\left(\bar{\mathbf{y}} - \mathbf{C}\mu\right)^T \mathbf{S}_{\mathbf{y}}^{-1}\left(\bar{\mathbf{y}} - \mathbf{C}\mu\right) \\ &= n\left(\bar{\mathbf{x}} - \mu\right)^T \mathbf{C}^T\left[\mathbf{C}\mathbf{S}_{\mathbf{x}}\mathbf{C}^T\right]^{-1}\mathbf{C}\left(\bar{\mathbf{x}} - \mu\right) \\ &\sim \frac{(n-1)\,q}{n-q}F_{n-q}^q. \qquad (7.1) \end{aligned}$$

To test $H_0 : \mathbf{C}\mu = \mathbf{0}$ we compute

$$\begin{aligned} T^2 &= n\bar{\mathbf{y}}^T\mathbf{S}_{\mathbf{y}}^{-1}\bar{\mathbf{y}} \\ &= n\bar{\mathbf{x}}^T\mathbf{C}^T\left[\mathbf{C}\mathbf{S}_{\mathbf{x}}\mathbf{C}^T\right]^{-1}\mathbf{C}\bar{\mathbf{x}}, \end{aligned}$$

whose null distribution is as at (7.1). This does not depend on *which* contrast matrix $\mathbf{C}$ we begin with, since any two contrast matrices $\mathbf{C}_1$, $\mathbf{C}_2$ have rows forming a basis for $\{\text{row}\,(1_p)\}^\perp$, hence are linear combinations of each other: $\mathbf{C}_1 = \mathbf{B}\mathbf{C}_2$ for some nonsingular $\mathbf{B}$, etc. – Asst. 1 #1(b).

- As in Example 6.2 we might however be interested in *particular* contrasts $\mathbf{c}_j^T$, thus defining a particular $\mathbf{C}$. We can work with $\mathbf{x}$ or $\mathbf{y}$, the latter seems easiest numerically. See R code on web site. The p-value associated with the hypothesis that $\mu_{\mathbf{y}} = \mathbf{0}$ is approximately 0, so we look to see which contrasts are significantly non-zero. The three simultaneous CIs on the three elements of $\mu_{\mathbf{y}}$ (which are the three contrasts $\mathbf{c}_j^T \mu_{\mathbf{x}}$) are

```
                 lower      ybar     upper
H effect     135.65030 209.31579 282.98128
C effect    -114.72708 -60.05263  -5.37818
interaction  -78.72858 -12.78947  53.14964
```

We conclude that there is no interaction effect, but both main effects are significant.

8. Comparing two population means; profile analysis

- Two independent samples:

$$\begin{array}{ll} \{\mathbf{x}_{1i}\}_{i=1}^{n_1} & \text{from a } N_p(\mu_1, \mathbf{\Sigma}_1) \text{ population,} \\ \{\mathbf{x}_{2i}\}_{i=1}^{n_2} & \text{from a } N_p(\mu_2, \mathbf{\Sigma}_2) \text{ population.} \end{array}$$

  We wish to make inferences about $\delta = \mu_1 - \mu_2$. (Recall Example 6.3 from the last class.)

- First assume that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ $(= \mathbf{\Sigma})$. Then the common value can be estimated, without bias, by the pooled covariance matrix (why?)

$$\mathbf{S}_{pooled} = \frac{(n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2}{n_1 + n_2 - 2}.$$

  The numerator is the sum of two independent Wishart matrices distributed as $W_p(n_1 - 1, \mathbf{\Sigma})$ and $W_p(n_2 - 1, \mathbf{\Sigma})$, hence is $\sim W_p(n_1 + n_2 - 2, \mathbf{\Sigma})$. Recall (5.1): for a hypothesized μ_0 we have that

$$(\bar{\mathbf{x}} - \mu_0)^T \left[\widehat{\text{cov}(\bar{\mathbf{x}})}\right]^{-1} (\bar{\mathbf{x}} - \mu_0) \sim dfS\frac{df1}{df2} F_{df2}^{df1}\left(\lambda^2\right).$$

In the current situation we replace $\bar{\mathbf{x}} - \mu_0$ by $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0$, whose covariance $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}$ is estimated by $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}$. Then (i) $dfS = n_1 + n_2 - 2$, (ii) $df1 = p$, (iii) $df2 = dfS + 1 - df1 = n_1 + n_2 - p - 1$. This gives

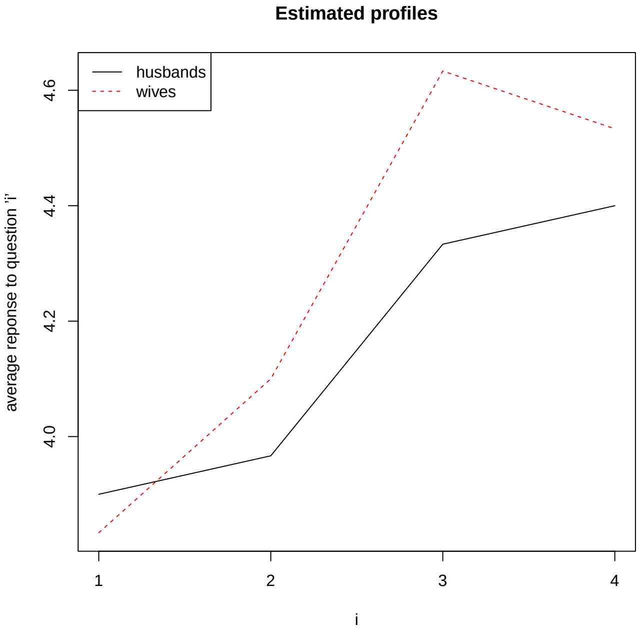
$$\begin{aligned} T^2 &= \left\{ \begin{array}{c} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0\right)^T \\ \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta_0\right) \end{array} \right\} \\ &\sim \frac{\left(n_1 + n_2 - 2\right)p}{n_1 + n_2 - 1 - p} F^{p}_{n_1+n_2-1-p}\left(\lambda^2\right), \text{ with} \qquad (8.1) \\ \lambda^2 &= \left(\delta - \delta_0\right)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}\right]^{-1} \left(\delta - \delta_0\right). \end{aligned}$$

Now inferences about δ can be made in the usual way. If $H_0 : \delta = \delta_0$ is true then the F is central.

- **Profile analysis**. If $\mathbf{x}$ is the result of p ‘treatments’, all measured in the same units – e.g. X_i a numerical response to the i^{th} question on a questionnaire – with mean vector $\mu_{p\times 1}$ then a line-graph of μ_i against i is the ‘profile’ for the population. Given two independent samples $\{\mathbf{x}_{1i}\}_{i=1}^{n_1}$

and $\{\mathbf{x}_{2i}\}_{i=1}^{n_2}$ with means μ_1 and μ_2, we seek to compare the profiles.

- Example – Problem 6.27, data in Table 6.14. Samples of married men and women (not married to each other – independent samples) rate their spouses with respect to $p = 4$ criteria by giving numerical responses (1-5) to questions. There are $n_1 = 30$ husbands and $n_2 = 30$ wives .



Estimated profiles: plots of $\bar{\mathbf{x}}_i$ against i for each group - husbands and wives.

1. Are the profiles parallel? This holds iff $\mu_{1,i} - \mu_{1,i-1} = \mu_{2,i} - \mu_{2,i-1}$ for $i = 2, ..., p$. So we test $H_{01} : \mathbf{C}\mu_1 = \mathbf{C}\mu_2$ for
$$\mathbf{C}_{p-1\times p} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$
Equivalently, transform to $\mathbf{y}_{li} = \mathbf{C}\mathbf{x}_{li}$, $l = 1, 2$, and test for equality of the two means – $\mu_{\mathbf{y}_1} = \mu_{\mathbf{y}_2}$ – using (8.1) with $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ replaced by $\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2$, δ by $\mathbf{0}$ and p by the dimension of each $\bar{\mathbf{y}}$.

2. If the profiles are parallel, are they coincident? Test $H_{02} : \mathbf{1}^T\mu_1 = \mathbf{1}^T\mu_2$ via a t-test based on the univariate samples $\left\{\mathbf{1}^T\mathbf{x}_{li}\right\}$, $l = 1, 2$. [Note $H_{01} + H_{02} \Rightarrow \mu_1 = \mu_2$.]

3. If the profiles are coincident, are they level? – i.e. we have decided that $\mu_{1,i} = \mu_{2,i}$ for $i = 1, 2, ..., p$. Is $\mu_{1,i} - \mu_{1,i-1} = \mu_{2,i} - \mu_{2,i-1} \overset{?}{=} 0$ for $i = 2, ..., p$? Combine the two samples into

one sample of size $n_1 + n_2$, with mean μ and covariance $\mathbf{\Sigma}$, and test $\mathbf{C}\mu = 0$. Equivalently, test that $\mu_{\mathbf{y}} = \mathbf{0}$, on the basis of the combined sample of $n_1 + n_2$ Y's.

- What if the covariance matrices cannot be assumed to be equal? If there is evidence that the two are not 'too far' apart (perhaps after transforming one of more of the variables) then the procedure above is still reasonably valid. Otherwise, note that

$$\text{cov}\left[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right] = \frac{1}{n_1}\mathbf{\Sigma}_1 + \frac{1}{n_2}\mathbf{\Sigma}_2$$

is estimated, without bias, by $\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2$, so that replacing $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}$ by this has intuitive appeal. The two are equal if the sample sizes are equal. The procedure is also asymptotically correct. But when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ $(= \mathbf{\Sigma})$,

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled} \sim \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{W_p\left(n_1 + n_2 - 2, \mathbf{\Sigma}\right)}{n_1 + n_2 - 2},$$

whereas now

$$\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2 \sim \frac{1}{n_1}\frac{W_p\left(n_1-1,\mathbf{\Sigma}_1\right)}{n_1-1}+\frac{1}{n_2}\frac{W_p\left(n_2-1,\mathbf{\Sigma}_2\right)}{n_2-1}.$$

As a result, the distribution of $\tilde{T}^2$ ($= T^2$ with the replacement described above) depends on both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ (even in the univariate case – 'Behrens-Fisher problem'). So the scaled F distribution is just an approximation, becoming more accurate (and $\stackrel{L}{\to} \chi^2_p$) as $n_1, n_2 \to \infty$.

- There is an approximation, which uses random df and seems reasonably accurate for normal populations. It involves replacing $dfS = n_1 + n_2 - 2$ by a random ν:

$$\begin{aligned}
\tilde{T}^2 &\sim \frac{\nu p}{\nu+1-p}F^p_{\nu+1-p}, \text{ where} \\
\nu &= \frac{p+p^2}{\sum_{i=1,2}\frac{1}{n_i}\left\{tr\mathbf{A}_i^2+\left(tr\mathbf{A}_i\right)^2\right\}} \text{ for} \\
\mathbf{A}_i &= \frac{\mathbf{S}_i}{n_i}\left(\frac{\mathbf{S}_1}{n_1}+\frac{\mathbf{S}_2}{n_2}\right)^{-1}.
\end{aligned}$$

- Husbands and wives example – see R code on course web site. Testing H_{01}, by transforming to $\mathbf{y}_{li} = \mathbf{C}\mathbf{x}_{li}$, $l = 1, 2$, and then either computing T^2, or using the manova() function in R, gives the p-value .06. Going on to test H_{02} by a univariate t-test gives $p = .22$; then H_{03} has $p = .00015$. Interpretation of these p-values?

- Dropping the assumption of equal covariances has almost no effect.

- What if these husbands and wives were married to each other?

9. Several populations: MANOVA

- Consider g normal populations, with means $\mu_1, ..., \mu_g$ and common covariance matrix Σ. Take a sample of size n_l from population l. We seek the unique effect of the l^{th} 'treatment' administered in population l; thus put

$$\begin{aligned} n &= \sum_l n_l, \text{ (total sample size)}, \\ \mu &= \sum_l \frac{n_l}{n} \mu_l, \text{ (weighted average)}, \\ \tau_l &= \mu_l - \mu, \text{ (unique effect)}. \end{aligned}$$

Then

$$\mu_l = \mu + \tau_l \text{ with } \sum_l n_l \tau_l = \mathbf{0}.$$

Rationale: we want parameters expressing the unique effects of the treatments, over and above the overall mean effect. One must be eliminated; the above is one way to accomplish this.

- **MLEs**: The log-likelihood is

$$\begin{aligned}\log L &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| \\ &\quad -\frac{1}{2}tr\left\{\boldsymbol{\Sigma}^{-1}\sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\mathbf{x}_{lj}-\mu_l\right)\left(\mathbf{x}_{lj}-\mu_l\right)^T\right\}.\end{aligned}$$

Denote

$$\begin{aligned}\bar{\mathbf{x}}_l &= \frac{1}{n_l}\sum_{j=1}^{n_l}\mathbf{x}_{lj}, \text{ so that} \\ \bar{\mathbf{x}} &= \frac{1}{n}\sum_{l=1}^{g}\sum_{j=1}^{n_l}\mathbf{x}_{lj} = \sum_l \frac{n_l}{n}\bar{\mathbf{x}}_l.\end{aligned}$$

Since $\mu_l = \mu + \tau_l$, the 'sum of squares and cross products' (SSCP) in L is

$$\begin{aligned}&= \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\mathbf{x}_{lj}-\tau_l-\mu\right)(\cdots)^T \\ &= \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l\right)+\left(\bar{\mathbf{x}}_l-\bar{\mathbf{x}}-\tau_l\right)+\left(\bar{\mathbf{x}}-\mu\right)\right)()^T\end{aligned}$$

$$\begin{aligned}
&= \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l\right)(\cdots)^T \\
&\quad +\sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\bar{\mathbf{x}}_l-\bar{\mathbf{x}}-\tau_l\right)(\cdots)^T \\
&\quad +\sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\bar{\mathbf{x}}-\mu\right)(\cdots)^T \text{ (why is this all?)} \\
&= \mathbf{W}+\sum_{l=1}^{g} n_l\left(\bar{\mathbf{x}}_l-\bar{\mathbf{x}}-\tau_l\right)(\cdots)^T + n\left(\bar{\mathbf{x}}-\mu\right)(\cdots)^T .
\end{aligned}$$

The trace in $\log L$ is minimized, hence L is maximized, by

$$\begin{aligned}
\hat{\tau}_l &= \bar{\mathbf{x}}_l-\bar{\mathbf{x}}, \\
\hat{\mu} &= \bar{\mathbf{x}},
\end{aligned}$$

so these are the MLEs. It remains to maximize

$$\log L_{|_{\hat{\tau}_l,\hat{\mu}}} = -\frac{np}{2}\log(2\pi)-\frac{n}{2}\log|\mathbf{\Sigma}|-\frac{1}{2}tr\mathbf{\Sigma}^{-1}\mathbf{W};$$

this results in the MLE $\hat{\mathbf{\Sigma}}=\mathbf{W}/n$ for

$$\begin{aligned}
\mathbf{W} &= \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l\right)\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l\right)^T \\
&= \sum_{l=1}^{g}\left(n_l-1\right)\mathbf{S}_l,
\end{aligned}$$

the 'within treatments' SSCP. Here $\mathbf{S}_l$ is the sample covariance in its group; the second form shows that

$$\mathbf{S}_{pooled} = \frac{1}{n-g}\mathbf{W}$$

is an unbiased estimate of $\boldsymbol{\Sigma}$. The maximized log-likelihood is

$$\max \log L = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\left|\frac{\mathbf{W}}{n}\right| - \frac{np}{2}.$$

- **Testing**: under the hypothesis that all treatments are equally effective, i.e. $H_0 : \tau_1 = \cdots = \tau_g$ $(= 0$ – why?) we have (after maximizing L over the remaining parameter μ)

$$\begin{aligned} \log L_{|_{\hat{\mu}}} &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| \\ &\quad -\frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\mathbf{W} - \frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\mathbf{B}, \end{aligned}$$

where

$$\mathbf{B} = \sum_{l=1}^{g} n_l\left(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_l - \bar{\mathbf{x}}\right)^T = \sum_{l=1}^{g} n_l\hat{\tau}_l\hat{\tau}_l^T$$

is the ‘between treatments’ SSCP. The restricted MLE of $\mathbf{\Sigma}$ is now $(\mathbf{B}+\mathbf{W})/n$, with

$$\max \log L_{|_{H_0}} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\left|\frac{\mathbf{B}+\mathbf{W}}{n}\right| - \frac{np}{2}.$$

Thus the LR test of H_0 rejects for small values of

$$\Lambda = \frac{\max L_{|_{H_0}}}{\max L} = \frac{\left|\frac{\mathbf{B}+\mathbf{W}}{n}\right|^{-\frac{n}{2}}}{\left|\frac{\mathbf{W}}{n}\right|^{-\frac{n}{2}}},$$

i.e. small values of “Wilks’ lambda”

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}.$$

Note that

$$\mathbf{B}+\mathbf{W} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}\right)^T.$$

- Results often expressed in a MANOVA table:

Source	SSCP	df
treatment	$\mathbf{B}$	$g-1$
error	$\mathbf{W}$	$n-g$
total	$\mathbf{B}+\mathbf{W}$	$n-1$

- We have that $\mathbf{W} \sim W_p\left(n-g, \boldsymbol{\Sigma}\right)$ and under H_0, $\mathbf{B} \sim W_p\left(g-1, \boldsymbol{\Sigma}\right)$ independently of $\mathbf{W}$ (how?), hence $\mathbf{B}+\mathbf{W} \sim W_p\left(n-1, \boldsymbol{\Sigma}\right)$.

- The exact (null) distributions of (transformations of) Λ^* are known in some special cases: $p = 1, 2$ and $g = 2, 3$ – see Table 6.3. Asymptotically, for large n we have Bartlett's correction
$$-\left(n-1-\frac{p+g}{2}\right)\log\Lambda^* \sim \chi^2_{p(g-1)}.$$
(The uncorrected form would be $-n\log\Lambda^* = -2\log\Lambda$.)

- Note that $\Lambda^* = \left|\left(\mathbf{B}+\mathbf{W}\right)^{-1}\mathbf{W}\right| = 1/\left|\mathbf{I}+\mathbf{W}^{-1}\mathbf{B}\right| = \prod\left(1+\lambda_i\right)^{-1}$, where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$. Other testing principles, e.g. union-intersection, lead to other tests, rejecting if the λ_i are in some overall sense large (discussion at p. 336):

  - Pillai – reject for large $tr\left[\left(\mathbf{B}+\mathbf{W}\right)^{-1}\mathbf{B}\right] = \sum\left(\lambda_i/\left(1+\lambda_i\right)\right)$.

  - Hotelling-Lawley – reject for large $tr\mathbf{W}^{-1}\mathbf{B} = \sum \lambda_i$.

  - Roy – reject for large $\max_i \lambda_i$ (union-intersection test).

- Example: Problem 6.24, data in Table 6.13 – $p =$ 4 measurements of various components of skulls from $g = 3$ time periods; $n_1 = n_2 = n_3 = 30$. R code on course web site.

- Test equality of covariances using Box's test – compares pooled with individual covariance estimates. Details assigned and in the text; Bartlett's correction gives $p = .39$. But, at least in the univariate case, if the group sizes are approximately equal then the anova tests are quite robust to heteroscedasticity and to non-normality. On the other hand the tests for variance/covariance equality can be very sensitive to non-normality.

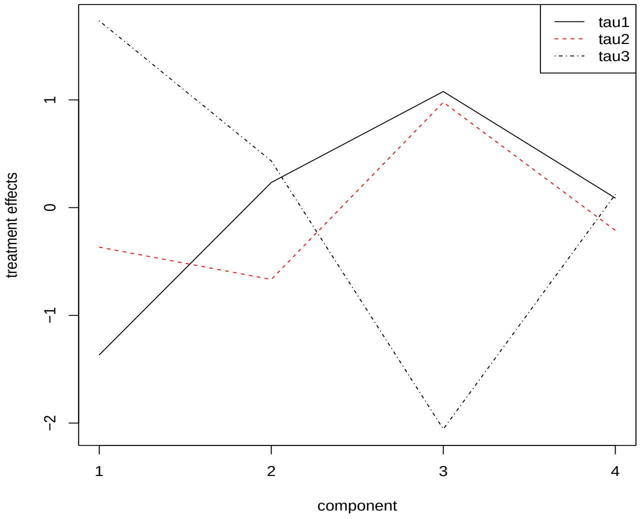
George Edward Pelham Box FRS (18 October 1919 – 28 March 2013)

‘To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!’

- Use manova() in R to test that all treatment effects are negligible: Wilks' Λ gives $p = .044$. Estimated effects:

$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$
-1.367	-0.367	1.733
0.233	-0.667	0.433
1.078	0.978	-2.056
0.089	-0.211	0.122

From the single aov's in the R output, and from the plot, the treatments are most different in the first and third components.



- To get CIs on marginal treatment effects: we have $\hat{\tau}_l = \bar{\mathbf{x}}_l - \bar{\mathbf{x}}$, so

  $$\hat{\tau}_k - \hat{\tau}_l = \bar{\mathbf{x}}_k - \bar{\mathbf{x}}_l \sim N_p\left(\tau_k - \tau_l, \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\mathbf{\Sigma}\right)$$

  and so, with $\hat{\mathbf{\Sigma}} = \mathbf{S}_{pooled} = \frac{1}{n-g}\mathbf{W}$ we have that

  $$\frac{(\hat{\tau}_{ki} - \hat{\tau}_{li}) - (\tau_{ki} - \tau_{li})}{\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_l}\right)\hat{\mathbf{\Sigma}}_{ii}}} \sim t_{n-g}.$$

  Then Bonferroni intervals on m such differences are

  $$\hat{\tau}_{ki} - \hat{\tau}_{li} \pm t_{n-g}\left(\frac{\alpha}{2m}\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_l}\right)\frac{w_{ii}}{n-g}}.$$

  There are $m = pg\left(g-1\right)/2$ such differences in total, with $l < k$. Even though the overall effects are significantly different at $\alpha = .012$, none of the individual differences are unless we go all the way to $\alpha = .25$:

```
75 % Bonferroni confidence intervals on
tau 2 minus tau 1 are:
     lower  point upper sig
[1,] -1.753 1     3.753
[2,] -3.758 -0.9  1.958
[3,] -3.123 -0.1  2.923
[4,] -2.189 -0.3  1.589

75 % Bonferroni confidence intervals on
tau 3 minus tau 1 are:
     lower  point  upper sig
[1,] 0.347  3.1    5.853   *
[2,] -2.658 0.2    3.058
[3,] -6.157 -3.133 -0.11   *
[4,] -1.855 0.033  1.922

75 % Bonferroni confidence intervals on
tau 3 minus tau 2 are:
     lower  point  upper sig
[1,] -0.653 2.1    4.853
[2,] -1.758 1.1    3.958
[3,] -6.057 -3.033 -0.01   *
[4,] -1.555 0.333  2.222
```

10. Two-way MANOVA; growth curves

- Consider gb normal populations classified two ways, so that the means represent effects of Factor 1, at g levels, and Factor 2, at b levels, and possible interactions between them. Each population has covariance matrix $\boldsymbol{\Sigma}$. Take a sample of size n from each. We seek the unique effects of the factors; thus the mean effects are represented as

$$\begin{aligned} \mu_{lk} &= \mu + \tau_l + \beta_k + \gamma_{lk}, \\ \text{with } \sum_{l=1}^{g} \tau_l &= \sum_{k=1}^{b} \beta_k = \sum_l \gamma_{kl} = \sum_k \gamma_{kl} = \mathbf{0}. \end{aligned}$$

This is one way of enforcing unique parameters, and is similar to what was done in one-way manova, when all cell sizes are equal (e.g. $\mu = \frac{1}{bg} \sum_{l,k} \mu_{lk}$, $\tau_l = \mu_{l\cdot} - \mu$, with $\mu_{l\cdot} = \frac{1}{b} \sum_k \mu_{lk}$, etc.). The 'unbalanced' case, in which the cell sizes might differ, is more problematic and is typically treated with regression methods (assigned).

- We decompose

$$\begin{aligned}
&\sum_{l,k,r} (\mathbf{x}_{lkr} - \mu_{lk})(\cdot\cdot)^T \\
= &\sum_{l,k,r} (\mathbf{x}_{lkr} - \mu - \tau_l - \beta_k - \gamma_{lk})(\cdot\cdot)^T \\
= &\sum_{l,k,r} \left[\begin{array}{c} (\mathbf{x}_{lkr} - \bar{\mathbf{x}}_{lk}) + \\ (\bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}} - \tau_l) + \\ (\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}} - \beta_k) + \\ (\bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}} - \gamma_{lk}) + \\ (\bar{\mathbf{x}} - \mu) \end{array} \right] [\cdot\cdot]^T
\end{aligned}$$

into the sum of SSCPs

$$\begin{aligned}
\sum_{l,k,r} (\mathbf{x}_{lkr} - \bar{\mathbf{x}}_{lk})(\cdot\cdot)^T &\quad + \\
\sum_{l,k,r} (\bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}} - \tau_l)(\cdot\cdot)^T &\quad + \\
\sum_{l,k,r} (\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}} - \beta_k)(\cdot\cdot)^T &\quad + \\
\sum_{l,k,r} (\bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}} - \gamma_{lk})(\cdot\cdot)^T &\quad + \\
\sum_{l,k,r} (\bar{\mathbf{x}} - \mu)(\cdot\cdot)^T &\quad ,
\end{aligned}$$

obtaining the MLEs

$$\begin{aligned}\hat{\tau}_l &= \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}},\\ \hat{\beta}_k &= \bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}},\\ \hat{\gamma}_{lk} &= \bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}},\\ \hat{\mu} &= \bar{\mathbf{x}},\end{aligned}$$

(hence $\hat{\mu}_{lk} = \bar{\mathbf{x}}_{lk}$) and the minimum SSCP

$$\sum_{lkr} \left(\bar{\mathbf{x}}_{lkr} - \bar{\mathbf{x}}_{lk}\right)(\cdot\cdot)^T \overset{def}{=} \mathbf{SSP}_{res}.$$

- The hypothesis that all τ_l vanish contributes a further

  $$\sum_{l,k,r} \left(\bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}\right)(\cdot\cdot)^T = bn \sum_{l} \left(\bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}\right)(\cdot\cdot)^T = \mathbf{SSP}_{fac1}.$$

  Similarly the hypotheses of no factor 2 effects, or no interactions, contribute

  $$\sum_{l,k,r} \left(\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}}\right)(\cdot\cdot)^T = gn \sum_{k} \left(\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}}\right)(\cdot\cdot)^T = \mathbf{SSP}_{fac2},$$

and

$$\begin{aligned}
& \sum_{l,k,r} \left(\bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}}\right)(\cdot\cdot)^T \\
= \; & n \sum_{l,k} \left(\bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}}\right)(\cdot\cdot)^T = \mathbf{SSP}_{int},
\end{aligned}$$

respectively.

- MANOVA table:

Source	SSCP	df
Factor 1	$\mathbf{SSP}_{fac1}$	$g-1$
Factor 2	$\mathbf{SSP}_{fac2}$	$b-1$
Interactions	$\mathbf{SSP}_{int}$	$(g-1)(b-1)$
error	$\mathbf{SSP}_{res}$	$gb(n-1)$
total		$gbn-1$

- The LR tests of H_{01} : no interactions, H_{02} : no Factor 1 effects, and H_{03} : no Factor 2 effects reject for small values of

$$\begin{aligned}\Lambda_1^* &= \frac{|\mathbf{SSP}_{res}|}{|\mathbf{SSP}_{int}+\mathbf{SSP}_{res}|},\\ \Lambda_2^* &= \frac{|\mathbf{SSP}_{res}|}{\left|\mathbf{SSP}_{fac1}+\mathbf{SSP}_{res}\right|},\\ \Lambda_3^* &= \frac{|\mathbf{SSP}_{res}|}{\left|\mathbf{SSP}_{fac2}+\mathbf{SSP}_{res}\right|},\end{aligned}$$

respectively. In each case there is a Bartlett's correction:

$$-\left(df_{res}-\frac{p+1-df_{effect}}{2}\right)\log\Lambda^*\sim\chi^2_{p\cdot df_{effect}}.$$

One would naturally test H_{01} first; if there are significant interactions then the other two hypotheses lose their meaning and one would probably revert to univariate anova's, seeking those individual responses in which interactions are, or are not, significant and then proceeding accordingly.

- **CIs on marginal differences of effects**. Suppose there are no significant interactions, so that it makes sense to compare the factor 1 effects. Then $\hat{\tau}_l - \hat{\tau}_m = \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{m\cdot} \sim N_p\left(\tau_l - \tau_m, \left(\frac{1}{bn} + \frac{1}{bn}\right)\mathbf{\Sigma}\right)$, and $\mathbf{E} \overset{def}{=} \mathbf{SSP}_{res} \sim W_p\left(df_{res}, \mathbf{\Sigma}\right)$. It follows that

  $$\frac{\left(\hat{\tau}_{l,i} - \hat{\tau}_{m,i}\right) - \left(\tau_{l,i} - \tau_{m,i}\right)}{\sqrt{\frac{2}{bn}\frac{E_{ii}}{df_{res}}}} \sim t_{df_{res}},$$

  from which CIs can be constructed. For Bonferroni intervals note that there are $pg\left(g-1\right)/2$ possible differences $\hat{\tau}_{l,i} - \hat{\tau}_{m,i}$ and $pb\left(b-1\right)/2$ possible differences $\hat{\beta}_{k,i} - \hat{\beta}_{q,i}$.

- **Growth curves**. A variation on repeated measures – treatment l $(l = 1, ..., g)$ is applied to each of n_l subjects and then a certain characteristic is monitored over time. For instance each of $n = \sum n_l$ plants is fertilized (using one of the g fertilizers) and their weights are measured at times $t_1, .., t_p$. Any one plant is viewed as an observation from a $N_p(\mu, \boldsymbol{\Sigma})$ population with $\mu_i =$ mean size at time t_i. The 'Potthoff-Roy model for quadratic growth' (there are others) takes

$$\mu = \begin{pmatrix} \beta_0 + \beta_1 t_1 + \beta_2 t_1^2 \\ \vdots \\ \beta_0 + \beta_1 t_p + \beta_2 t_p^2 \end{pmatrix} = \mathbf{B}\beta,$$

in an obvious notation. The coefficient vectors β vary from one group to another. Thus the sample data are

$$\left\{ \mathbf{x}_{lj} \,|\, j = 1, ..., n_l, l = 1, ..., g \right\}$$

with $\mathbf{x}_{lj} \sim N_p(\mu_l = \mathbf{B}\beta_l, \boldsymbol{\Sigma})$. We wish to compare the curves in varying groups. In the unrestricted model, with no assumed structure on μ_l, the mles are $\hat{\mu}_l = \bar{\mathbf{x}}_l$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{W}/n$, where $\mathbf{W} = (n-g)\,\mathbf{S}_{pooled}$, exactly as in one-way manova. To test the adequacy of a particular growth curve model $\mu_l = \mathbf{B}\beta_l$, where $\mathbf{B}$ is $p \times (q+1)$ (for instance when a q^{th}-order polynomial is fitted), we must find the mle's of the β_l. These minimize the trace in the exponent of the likelihood, which is:

$$\begin{aligned} & tr\boldsymbol{\Sigma}^{-1}\left\{\sum_l\sum_j\left(\mathbf{x}_{lj}-\mathbf{B}\beta_l\right)(\cdots)^T\right\} \\ = \; & tr\boldsymbol{\Sigma}^{-1}\left\{\sum_l\sum_j\left(\left(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l\right)+\left(\bar{\mathbf{x}}_l-\mathbf{B}\beta_l\right)\right)(\cdots)^T\right\} \\ = \; & tr\boldsymbol{\Sigma}^{-1}\mathbf{W}+\sum_l n_l\left(\bar{\mathbf{x}}_l-\mathbf{B}\beta_l\right)^T\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}_l-\mathbf{B}\beta_l\right) \\ = \; & tr\boldsymbol{\Sigma}^{-1}\mathbf{W}+\sum_l n_l\left\|\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}}_l-\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\beta_l\right\|^2. \end{aligned}$$

Thus $\hat{\beta}_l$ minimizes $\left\|\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{x}}_l-\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\beta_l\right\|^2$; by standard least squares theory this is

$$\hat{\beta}_l=\left(\mathbf{B}^T\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_l. \tag{10.1}$$

Now Σ is replaced by the mle $\hat{\Sigma}$ in this model. To obtain this mle one would substitute $\hat{\beta}_l$ back into the likelihood and maximize (numerically) over Σ. It is instead usual (because it is much simpler) to instead take $\hat{\Sigma}=\mathbf{S}_{pooled}$, i.e. to use the mle from the ‘unrestricted’ model. The $\hat{\beta}_l$ are independent as l varies over subjects (why?). To test a particular growth model, i.e. $H_0 : \mu = \mathbf{B}\beta$ (for any fixed $\mathbf{B}$ and $\beta_{(q+1)\times 1}$, not necessarily representing quadratic effects) we fit without restrictions and then under the hypothesis, obtaining

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{W}_q|},$$

where

$$\begin{aligned} \mathbf{W}_q &= \sum_{l=1}^{g} \sum_{j=1}^{n_l} \left(\mathbf{x}_{lj} - \mathbf{B}\hat{\beta}_l\right) \left(\mathbf{x}_{lj} - \mathbf{B}\hat{\beta}_l\right)^T \\ &= \mathbf{W} + \sum_l n_l \left(\bar{\mathbf{x}}_l - \mathbf{B}\hat{\beta}_l\right) \left(\bar{\mathbf{x}}_l - \mathbf{B}\hat{\beta}_l\right)^T . \end{aligned}$$

Bartlett's approximation:

$$-\left(n - \frac{p-q+g}{2}\right) \log \Lambda^* \sim \chi^2_{(p-q-1)g}.$$

11. Multivariate regression: estimates and distributions

- Example 7.10 in text: An investigator seeks to assess the relationship between two types of computing hardware requirements – $Y_1 =$ CPU time and $Y_2 =$ input/output capacity – and the predictors $Z_1 =$ customer orders and $Z_2 =$ item count.

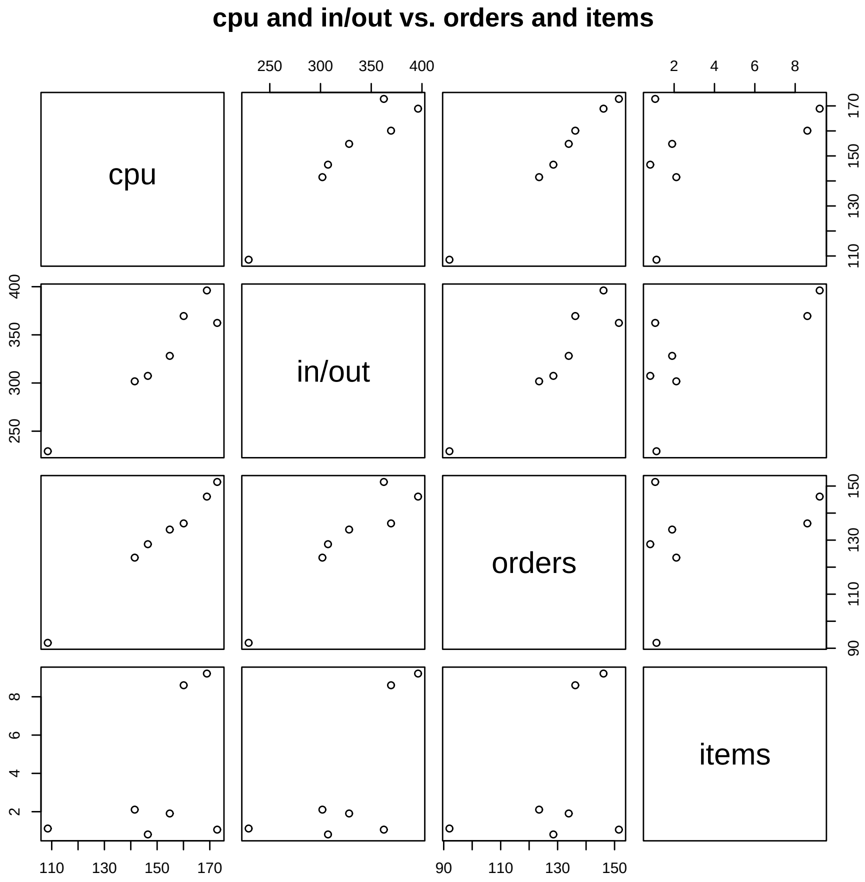
- See pairs() plot; R code on web site – some linear dependence on orders (at least). To assess the joint linear effect of orders and item count we could consider two linear regressions on $Z = (Z_1, Z_2)$ – one for each dependent variable. Given n observations on each we would fit

$$\mathbf{Y}_j = \mathbf{Z}\beta_j + \varepsilon,$$

where $\mathbf{Z}$ is the $n \times 3$ matrix $(\mathbf{1}_n, \mathbf{Z}_1, \mathbf{Z}_2)$, and $\mathbf{Y}_j$ $(j = 1, 2)$ contains the data on Y_j. One commonly assumes that $\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$, then the mle of β_j is the LSE:

$$\hat{\beta}_j = \arg\min \left\|\mathbf{Y}_j - \mathbf{Z}\beta\right\|^2 = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y}_j, \tag{11.1}$$

the mle of σ^2 is $\hat{\sigma}^2 = \left\| \mathbf{Y}_j - \mathbf{Z}\hat{\beta}_j \right\|^2 / n$ (but is biased; $S^2 = \left\| \mathbf{Y}_j - \mathbf{Z}\hat{\beta}_j \right\|^2 / (n - r - 1)$, where $r+1$ is the number of columns of $\mathbf{Z}$, is unbiased).



- The marginal approach above ignores possible dependencies between Y_1 and Y_2 (evident from the pairs plot). To account for this consider the multivariate regression model (with m responses $Y_1, ..., Y_m$ and r predictors $Z_1, ..., Z_r$). Define

$$\begin{aligned} \mathbf{Y}_{n\times m} &= \left(\mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_m\right), \\ \mathbf{Z}_{n\times(r+1)} &= \left(\mathbf{1}_n \vdots \mathbf{Z}_1 \vdots \cdots \vdots \mathbf{Z}_r\right), \\ \mathbf{B}_{(r+1)\times m} &= \left(\beta_1 \vdots \cdots \vdots \beta_m\right) \end{aligned}$$

and consider the model

$$\mathbf{Y} = \mathbf{ZB} + \mathbf{E},$$

where the <u>rows</u> $\boldsymbol{\varepsilon}_{(j)}^T$ of $\mathbf{E}_{n\times m}$ are i.i.d. $N_m(\mathbf{0}, \mathbf{\Sigma})$ r.vecs. At this point $\mathbf{Z}$ is viewed as fixed (not random). Thus if $\mathbf{Y}_{(j)}^T$ is the j^{th} row of $\mathbf{Y}$ and $\mathbf{z}_j^T$ is the j^{th} row of $\mathbf{Z}$, then $E\left[\mathbf{Y}_{(j)}^T\right] = \mathbf{z}_j^T\mathbf{B}$ and so $\mathbf{Y}_{(j)} \sim N_m\left(\mathbf{B}^T\mathbf{z}_j, \mathbf{\Sigma}\right)$.

- Kronecker product:

$$\mathbf{A}_{m\times n}\otimes\mathbf{B}_{p\times q}=\left(a_{ij}\mathbf{B}\right)\ : mp\times nq.$$

- Let $vec\,(\mathrm{M})$ consist of the *rows* of M stretched out as a long column. Then

$$\begin{aligned} vec\mathbf{Y} &= \begin{pmatrix}\mathbf{Y}_{(1)}\\ \vdots \\ \mathbf{Y}_{(n)}\end{pmatrix} = vec\,(\mathbf{ZB}) + vec\,(\mathbf{E})\,,\ \text{with} \\ vec\,(\mathbf{E}) &= \begin{pmatrix}\boldsymbol{\varepsilon}_{(1)}\\ \vdots \\ \boldsymbol{\varepsilon}_{(n)}\end{pmatrix} \sim N_{mn}\left(\mathbf{0},\mathbf{I}_n\otimes\mathbf{\Sigma}\right). \end{aligned}$$

Thus

$$vec\mathbf{Y}\sim N_{mn}\left(vec\,(\mathbf{ZB})\,,\mathbf{I}_n\otimes\mathbf{\Sigma}\right).$$

- Useful identities:

$$\begin{aligned} vec\,(\mathbf{ABC}) &= \left(\mathbf{A}\otimes\mathbf{C}^T\right)vec\mathbf{B}, \\ (\Leftarrow vec\left(\mathbf{xy}^T\right) &= \mathbf{x}\otimes\mathbf{y}) \\ (\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) &= (\mathbf{AC}\otimes\mathbf{BD})\,, \\ (vec\mathbf{A})^T\,vec\mathbf{B} &= tr\mathbf{AB}^T = tr\mathbf{B}^T\mathbf{A}. \end{aligned}$$

- The likelihood is

$$
\begin{aligned}
L(\mathbf{B},\mathbf{\Sigma}) = (2\pi)^{-\frac{mn}{2}} \left|\mathbf{I}_n \otimes \mathbf{\Sigma}\right|^{-1/2} \cdot \\
\exp\left\{ \begin{array}{c} -\frac{1}{2}\left(vec\left(\mathbf{Y}-\mathbf{ZB}\right)\right)^T \cdot \\ \left(\mathbf{I}_n \otimes \mathbf{\Sigma}\right)^{-1}\left(vec\left(\mathbf{Y}-\mathbf{ZB}\right)\right) \end{array} \right\} \\
= \quad & (2\pi)^{-\frac{mn}{2}} \left|\mathbf{\Sigma}\right|^{-\frac{n}{2}} \cdot \\
& \exp\left\{-\frac{1}{2}\left(vec\left(\mathbf{Y}-\mathbf{ZB}\right)\right)^T\left(vec\left(\mathbf{Y}-\mathbf{ZB}\right)\mathbf{\Sigma}^{-1}\right)\right\} \\
= \quad & (2\pi)^{-\frac{mn}{2}} \left|\mathbf{\Sigma}\right|^{-\frac{n}{2}} tr\mathbf{\Sigma}^{-1}\left(\mathbf{Y}-\mathbf{ZB}\right)^T\left(\mathbf{Y}-\mathbf{ZB}\right),
\end{aligned}
$$

with log-likelihood (apart from the constant)

$$
l(\mathbf{B},\mathbf{\Sigma}) = -\frac{n}{2}\log\left|\mathbf{\Sigma}\right| - \frac{1}{2}tr\mathbf{\Sigma}^{-1}\left(\mathbf{Y}-\mathbf{ZB}\right)^T\left(\mathbf{Y}-\mathbf{ZB}\right).
$$

Put $\mathbf{H}_{n\times n} = \mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T$. Properties of this important matrix:

- $\mathbf{HZ} = \mathbf{Z}$; hence $\mathbf{H}$ is idempotent (hence so is $\mathbf{I}-\mathbf{H}$, and $\left(\mathbf{I}-\mathbf{H}\right)\mathbf{H} = \mathbf{0}$.

- If $\mathbf{v}$ is an eigenvector and λ and eigenvalue – $\mathbf{Hv} = \lambda\mathbf{v}$ – then $\lambda \in \{0,1\}$, hence rank $=$ trace $= r+1$.

- The spectral decomposition $\mathbf{H} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$ has $\mathbf{D} = diag\left(\mathbf{I}_{r+1}, \mathbf{0}\right)$, hence $\mathbf{H} = \mathbf{V}_1\mathbf{V}_1^T$, where $\mathbf{V}_1$ consists of the first $r+1$ columns of $\mathbf{Q}$; since these are mutually orthogonal we have $\mathbf{V}_1^T\mathbf{V}_1 = \mathbf{I}_{r+1}$. Then $\mathbf{I} - \mathbf{H} = \mathbf{V}_2\mathbf{V}_2^T$, where $\mathbf{V}_2$ consists of the last $n-r-1$ columns of $\mathbf{Q}$.

- The diagonal elements of a symmetric matrix lie between the smallest and largest eigenvalues, hence $0 \leq h_{ii} \leq 1$.

- Proceed as in univariate regression: write

$$\begin{aligned} \mathbf{Y} - \mathbf{ZB} &= \left(\mathbf{I}_n - \mathbf{H}\right)\left(\mathbf{Y} - \mathbf{ZB}\right) + \mathbf{H}\left(\mathbf{Y} - \mathbf{ZB}\right) \\ &= \left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y} + \left(\mathbf{HY} - \mathbf{ZB}\right), \end{aligned}$$

where these two matrices are mutually orthogonal. Thus

$$\begin{aligned} l\left(\mathbf{B}, \boldsymbol{\Sigma}\right) &= -\frac{n}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}tr\boldsymbol{\Sigma}^{-1}\left\{\mathbf{Y}^T\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y}\right\} \\ &\quad -\frac{1}{2}tr\left(\mathbf{HY} - \mathbf{ZB}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{HY} - \mathbf{ZB}\right)^T. \end{aligned}$$

The second trace is non-negative (why?), and $= 0$ iff

$$\mathbf{B} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y} \overset{def}{=} \hat{\mathbf{B}};$$

thus in particular $\hat{\beta}_j$ is given by (11.1).

- With

$$\mathbf{W} = \mathbf{Y}^T\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y},$$

the partially maximized log-likelihood is

$$l\left(\hat{\mathbf{B}}, \mathbf{\Sigma}\right) = -\frac{n}{2}\log\left|\mathbf{\Sigma}\right| - \frac{1}{2}tr\mathbf{\Sigma}^{-1}\mathbf{W},$$

which as usual is maximized by

$$\hat{\mathbf{\Sigma}} = \mathbf{W}/n,$$

with

$$l\left(\hat{\mathbf{B}}, \hat{\mathbf{\Sigma}}\right) = const. - \frac{n}{2}\log\left|\hat{\mathbf{\Sigma}}\right| - \frac{nm}{2}.$$

Note that

$$\mathbf{W} = \left(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}\right)^T\left(\mathbf{Y} - \mathbf{Z}\hat{\mathbf{B}}\right) \overset{def}{=} \hat{\mathbf{E}}^T\hat{\mathbf{E}}.$$

- To get the distributions, write
$$\mathbf{I}_n - \mathbf{H} = \mathbf{V}_2 \mathbf{V}_2^T,$$
as above. Put
$$\mathbf{X}_{n-r-1\times m} = \mathbf{V}_2^T \mathbf{Y},$$
then
$$\mathbf{W} = \mathbf{X}^T \mathbf{X} = \sum_{j=1}^{n-r-1} \mathbf{x}_j \mathbf{x}_j^T,$$
where the $\mathbf{x}_j^T$ are the rows of $\mathbf{X}$.

- These are independent $N_m(\mathbf{0}, \mathbf{\Sigma})$ r.vecs, so that
$$\mathbf{W} \sim W_m(n-r-1, \mathbf{\Sigma})$$
and so $\mathbf{W}/(n-r-1)$ is an unbiased for $\mathbf{\Sigma}$.

Reason:

$$\begin{aligned}
vec\mathbf{X} &= vec\left(\mathbf{V}_2^T\mathbf{Y}\mathbf{I}_m\right) \\
&= \left(\mathbf{V}_2^T\otimes\mathbf{I}_m\right)vec\mathbf{Y} \\
&\sim N_q\left(\begin{array}{c}\left(\mathbf{V}_2^T\otimes\mathbf{I}_m\right)vec\left(\mathbf{ZB}\right), \\ \left(\mathbf{V}_2^T\otimes\mathbf{I}_m\right)\left(\mathbf{I}_n\otimes\boldsymbol{\Sigma}\right)\left(\mathbf{V}_2^T\otimes\mathbf{I}_m\right)^T\end{array}\right) \\
&= N_q\left(vec\left(\mathbf{V}_2^T\mathbf{ZB}\right),\mathbf{V}_2^T\mathbf{V}_2\otimes\boldsymbol{\Sigma}\right) \\
&= N_q\left(\mathbf{0},\mathbf{I}_{n-r-1}\otimes\boldsymbol{\Sigma}_{m\times m}\right),
\end{aligned}$$

(why $=\mathbf{0}$?) as required. (Here $q=m\left(n-r-1\right)$.)

12. Multivariate regression: inferences

- Since $\hat{\mathrm{B}}$ and $\mathbf{X}$ are linear functions of $\mathbf{Y}_{n\times m}$ they are jointly normally distributed. The mean of $\hat{\mathbf{B}}_{(r+1)\times m} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Y}$ is $\mathbf{B}$:

$$\begin{aligned} E\left[\hat{\mathbf{B}}\right] &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T E\left[\mathbf{Y}\right] \\ &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{B} \\ &= \mathbf{B}, \end{aligned}$$

and the covariance is

$$\begin{aligned} \mathrm{COV}\left[vec\hat{\mathbf{B}}\right] &= \left\{ \begin{array}{c} \left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right)\mathrm{COV}\left[vec\mathbf{Y}\right] \\ \cdot\left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right)^T \end{array} \right\} \\ &= \left\{ \begin{array}{c} \left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right)\left(\mathbf{I}_n \otimes \mathbf{\Sigma}\right) \\ \cdot\left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right)^T \end{array} \right\} \\ &= \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \otimes \mathbf{\Sigma}. \end{aligned}$$

The covariances between the elements of $\hat{\mathrm{B}}$ and those of $\mathbf{X}$ are the elements of

$$\begin{aligned}
& \text{COV}\left[vec\hat{\mathbf{B}}, (vec\mathbf{X})^T\right] \\
= \;& \text{COV}\left[\begin{array}{c} \left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right) vec\mathbf{Y}, \\ \left(\left(\mathbf{V}_2^T \otimes \mathbf{I}_m\right) vec\mathbf{Y}\right)^T \end{array}\right] \\
= \;& \left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right) \text{COV}\left[vec\mathbf{Y}\right]\left(\mathbf{V}_2 \otimes \mathbf{I}_m\right) \\
= \;& \left(\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T \otimes \mathbf{I}_m\right)\left(\mathbf{I}_n \otimes \mathbf{\Sigma}\right)\left(\mathbf{V}_2 \otimes \mathbf{I}_m\right) \\
= \;& \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{V}_2 \otimes \mathbf{\Sigma} \\
= \;& \mathbf{0}_{(r+1)\times(n-r-1)} \otimes \mathbf{\Sigma} \\
= \;& \mathbf{0}_{m(r+1)\times m(n-r-1)}.
\end{aligned}$$

Thus $\hat{\mathrm{B}}$ and $\mathbf{X}$ are jointly normal and uncorrelated, hence independent. Since $\mathbf{W}$ is a function of $\mathbf{X}$ alone, it too is independent of $\hat{\mathrm{B}}$.

- **Summary**:

$$vec\hat{\mathbf{B}} \sim N_{(r+1)m}\left(vec\mathbf{B}, \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \otimes \mathbf{\Sigma}\right),$$

independently of $\mathbf{W} \sim W_m\left(n-r-1, \mathbf{\Sigma}\right)$.

12.1. Testing

- Example 1: Partition $\mathbf{B}$ as
$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{(1)} \\ \mathbf{B}_{(2)} \end{pmatrix} \begin{array}{l} \leftarrow (s+1) \times m \\ \leftarrow (r-s) \times m \end{array}.$$
We might test that $\mathbf{B}_{(2)} = \mathbf{0}$, i.e. that none of the m mean responses depend on any of the last $r-s$ covariates (columns of $\mathbf{Z}$). Equivalently, $\mathbf{AB} = \mathbf{0}$, where
$$\mathbf{A}_{(r-s)\times(r+1)} = \left(\mathbf{0}_{(r-s)\times(s+1)} \vdots \mathbf{I}_{r-s}\right).$$

- Example 2: Suppose we test that, apart from the intercepts, all mean responses are equal; i.e.
$$\mathbf{B} = \begin{pmatrix} \beta_{1,0} & \cdots & \beta_{m,0} \\ \beta_{1,1} & \cdots & \beta_{m,1} \end{pmatrix} \begin{array}{l} \leftarrow 1 \times m \\ \leftarrow r \times m \end{array}$$
and $\beta_{1,1} = \cdots = \beta_{m,1}$. Equivalently,
$$(\mathbf{0} \vdots \mathbf{I}_r)\, \mathbf{B} \begin{pmatrix} -1 & & & \mathbf{0} \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ \mathbf{0} & & & 1 \end{pmatrix}_{m\times(m-1)} = \mathbf{0}_{r\times(m-1)}.$$

- Each is an instance of a 'General Linear Hypothesis'. In general form:

$$\mathbf{A}_{q\times(r+1)}\mathbf{B}_{(r+1)\times m}\mathbf{C}_{m\times v} = \mathbf{D}_{q\times v}.$$

  Here it is assumed that the rows of $\mathbf{A}$ are independent, so that $rk\,(\mathbf{A}) = q \leq r+1$, and that the columns of $\mathbf{C}$ are independent, so that $rk\,(\mathbf{C}) = v \leq m$. (Commonly $\mathbf{C} = \mathbf{I}_m$ or $\mathbf{A} = \mathbf{I}_q$.)

- First define

$$\tilde{Y}_{n\times v} = \mathbf{YC} = \mathbf{ZBC} + \mathbf{EC} = \mathbf{Z}\tilde{B} + \tilde{E},$$

  where

$$\begin{aligned} \tilde{B}_{(r+1)\times v} &= \mathbf{BC}, \\ \tilde{E}_{n\times v} &= \mathbf{EC}, \end{aligned}$$

  and

$$vec\tilde{E} \sim N_{vn}\left(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{C}^T\mathbf{\Sigma}\mathbf{C}\right) = N_{vn}\left(\mathbf{0}, \mathbf{I}_n \otimes \tilde{\mathbf{\Sigma}}\right),$$

  for $\tilde{\mathbf{\Sigma}}_{v\times v} = \mathbf{C}^T\mathbf{\Sigma}\mathbf{C} > \mathbf{0}$. In this notation,

$$vec\tilde{Y} \sim N_{vn}\left(vec\mathbf{Z}\tilde{B}, \mathbf{I}_n \otimes \tilde{\mathbf{\Sigma}}\right),$$

and we test that $\mathbf{A}_{q\times(r+1)}\tilde{B}_{(r+1)\times v} = \mathbf{D}_{q\times v}$. The regression estimates (mles) are

$$\widehat{\tilde{B}} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\tilde{Y} = \hat{\mathbf{B}}\mathbf{C},$$

with

$$vec\widehat{\tilde{B}} \sim N_{(r+1)v}\left(vec\tilde{B}, \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\otimes\tilde{\mathbf{\Sigma}}\right)$$

and

$$\begin{aligned} vec\mathbf{A}\widehat{\tilde{B}} &= \left(\mathbf{A}\otimes\mathbf{I}_v\right)vec\widehat{\tilde{B}} \\ &\sim N_{qv}\left(vec\mathbf{A}\tilde{B}, \mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T\otimes\tilde{\mathbf{\Sigma}}\right). \end{aligned}$$

Thus, under the hypothesis,

$$vec\left(\mathbf{A}\widehat{\tilde{B}}-\mathbf{D}\right) \sim N_{qv}\left(\mathbf{0}, \mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T\otimes\tilde{\mathbf{\Sigma}}\right)$$

and

$$\begin{aligned} & vec\left(\left[\mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T\right]^{-1/2}\left(\mathbf{A}\widehat{\tilde{B}}-\mathbf{D}\right)\right) \\ \sim\ & N_{qv}\left(\mathbf{0}, \mathbf{I}_q\otimes\tilde{\mathbf{\Sigma}}\right). \end{aligned}$$

This means that the rows $\mathbf{x}_1^T, ..., \mathbf{x}_q^T$ of $\left[\mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T\right]^{-1/2}\left(\mathbf{A}\widehat{\tilde{B}}-\mathbf{D}\right)$ are i.i.d. $N_v\left(\mathbf{0},\tilde{\mathbf{\Sigma}}\right)$, and so

$$\begin{aligned}
\mathbf{R}_{v\times v} &\overset{def}{=} \sum_{i=1}^{q} \mathbf{x}_i \mathbf{x}_i^T \\
&= \left(\mathbf{A}\widehat{\tilde{B}} - \mathbf{D}\right)^T \left[\mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T\right]^{-1}\left(\mathbf{A}\widehat{\tilde{B}} - \mathbf{D}\right) \\
&= \left(\mathbf{A}\hat{\mathbf{B}}\mathbf{C} - \mathbf{D}\right)^T \left[\mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T\right]^{-1}\left(\mathbf{A}\hat{\mathbf{B}}\mathbf{C} - \mathbf{D}\right) \\
&\sim W_v\left(q, \tilde{\mathbf{\Sigma}}\right) = W_v\left(q, \mathbf{C}^T\mathbf{\Sigma}\mathbf{C}\right),
\end{aligned}$$

independently of $\tilde{W} \sim W_v\left(n - r - 1, \mathbf{C}^T\mathbf{\Sigma}\mathbf{C}\right)$. ($\tilde{W}$ is computed from $\tilde{Y}$, and $= \mathbf{C}^T\mathbf{W}\mathbf{C}$.)

- Now inferences are based on the eigenvalues of $\tilde{W}^{-1}\mathbf{R}$, as in MANOVA. For instance the LR test rejects for small values of

$$\Lambda^* = \frac{\left|\tilde{W}\right|}{\left|\mathbf{R} + \tilde{W}\right|}.$$

There is a built-in R function to carry out this and various other tests; the user inputs only $\mathbf{A}$, $\mathbf{C}$ and $\mathbf{D}$.

- Hardware example – code on web site.
  - Test overall regression, i.e. that the last 2 of the 3 rows of $\mathbf{B}$ vanish:

$$\mathbf{B}_{(2)} = \mathbf{ABC} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{B} = \mathbf{0}.$$

12.2. Intervals

- A summary of the results above is that, with $\mathbf{A}_{q\times(r+1)}$ and $\mathbf{C}_{m\times v}$,

$$\begin{aligned} & vec\left(\mathbf{A}\hat{\mathbf{B}}\mathbf{C}\right) \\ \sim \; & N_{qv}\left(vec\left(\mathbf{ABC}\right), \mathbf{A}\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{A}^T \otimes \mathbf{C}^T\mathbf{\Sigma}\mathbf{C}\right). \end{aligned}$$

The vector of mean responses at a particular level $\mathbf{z}_*$ is $E\left[(Y_1, ..., Y_m)\,|\mathbf{z}_*\right] = \mathbf{z}_*^T\mathbf{B} \overset{def}{=} \gamma_*^T$ and is estimated by $\hat{\gamma}_*^T = \mathbf{z}_*^T\hat{\mathbf{B}}$, with

$$\begin{aligned} \hat{\gamma}_* &= vec\left(\mathbf{z}_*^T\hat{\mathbf{B}}\right) \\ &\sim N_m\left(vec\left(\mathbf{z}_*^T\mathbf{B}\right), \mathbf{z}_*^T\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{z}_* \otimes \mathbf{\Sigma}\right) \\ &= N_m\left(\gamma_*, \left[\mathbf{z}_*^T\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{z}_*\right] \cdot \mathbf{\Sigma}\right). \end{aligned}$$

Thus $(\hat{\gamma}_* - \gamma_*)/\sqrt{\mathbf{z}_*^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}_*} \sim N_m(\mathbf{0}, \mathbf{\Sigma})$; inserting $\hat{\mathbf{\Sigma}} = \mathbf{W}/(n-r-1)$ gives

$$\begin{aligned} T^2 &= \frac{(\hat{\gamma}_* - \gamma_*)^T \left[\frac{\mathbf{W}}{n-r-1}\right]^{-1} (\hat{\gamma}_* - \gamma_*)}{\mathbf{z}_*^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}_*} \\ &\sim dfS \frac{df1}{df2} F_{df2}^{df1} = (n-r-1)\frac{m}{n-r-m} F_{n-r-m}^m. \end{aligned}$$

(The Wishart distribution associated with the unbiased covariance estimate is $W_{df1}(dfS, \cdot)$ and $df1 + df2 = dfS + 1$.) Thus a $100(1-\alpha)\%$ confidence ellipsoid is

$$\left\{ \begin{array}{c} \gamma_* | \frac{(\hat{\gamma}_* - \gamma_*)^T \left[\frac{\mathbf{W}}{n-r-1}\right]^{-1} (\hat{\gamma}_* - \gamma_*)}{\mathbf{z}_*^T (\mathbf{Z}^T\mathbf{Z})^{-1} \mathbf{z}_*} \\ \leq \frac{m(n-r-1)}{n-r-m} F_{n-r-m}^m(\alpha) \overset{def}{=} c^2 \end{array} \right\}.$$

- A $100(1-\alpha)\%$ prediction region for a *new* random vector at this level – $\mathbf{y}_{new} \sim N_m\left(\mathbf{B}^T\mathbf{z}_* = \gamma_*, \mathbf{\Sigma}\right)$ – is obtained as follows. The prediction is $\hat{\mathbf{B}}^T\mathbf{z}_* = \hat{\gamma}_* \sim N_m\left(\gamma_*, \left[\mathbf{z}_*^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}_*\right] \cdot \mathbf{\Sigma}\right)$ as above, and this is independent of $\mathbf{y}_{new}$ (why?) and then

$$\mathbf{y}_{new} - \hat{\gamma}_* \sim N_m\left(0, \left[1 + \mathbf{z}_*^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}_*\right] \cdot \mathbf{\Sigma}\right);$$

thus the prediction region is

$$\left\{\mathbf{y}_{new} | \frac{(\mathbf{y}_{new} - \hat{\gamma}_*)^T \left[\frac{\mathbf{W}}{n-r-1}\right]^{-1} (\mathbf{y}_{new} - \hat{\gamma}_*)}{1 + \mathbf{z}_*^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}_*} \le c^2\right\}.$$

- Hardware example – Confidence and prediction regions when $\mathbf{z}_0^T = (1, 130, 7.5)$ – code on web site.
  - $\frac{\mathbf{W}}{n-r-1} = \mathbf{U}^T\mathbf{U}$, $t^2 = \mathbf{z}_*^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}_* \cdot c^2$ for the confidence region,
  - $\mathbf{v} = \mathbf{U}^{-T}(\hat{\gamma}_* - \gamma_*)$, $\|\mathbf{v}\|^2 \le t^2$ describes the region,
  - plot $\gamma_* = \hat{\gamma}_* - \mathbf{U}^T\mathbf{v}$ as $\mathbf{v}$ ranges over the boundary of this sphere of radius t.

13. General concepts of linear regression

- **Best mse prediction**: In Lectures 11, 12 the regressors were viewed as *fixed*. What if they, along with Y, are *random*? Given *any* r.v.s $(Y, \mathbf{z})$, the function $h(\mathbf{z})$ minimizing the mean squared prediction error $E\left[\{Y - h(\mathbf{z})\}^2\right]$ is $h(\mathbf{z}) = E[Y|\mathbf{z}]$. (See STAT 479 Lecture 8 if this is unfamiliar to you.)

- As in Lecture 3, if the joint distribution of $(Y, \mathbf{z})$ is *normal*, then
$$E[Y|\mathbf{z}] = \mu_Y + \sigma_{Y\mathbf{z}}^T \mathbf{\Sigma}_{\mathbf{zz}}^{-1}(\mathbf{z} - \mu_{\mathbf{z}})$$
is *linear* in $\mathbf{z}$, i.e. the best (minimum mse) predictor in this case is linear.

- Even if the joint distribution is non-normal, the best predictor of Y which is a *linear* function $\beta_0 + \beta_1^T \mathbf{z}$ is given by
$$\begin{aligned} \beta_0 &= \mu_Y - \sigma_{Y\mathbf{z}}^T \mathbf{\Sigma}_{\mathbf{zz}}^{-1} \mu_{\mathbf{z}}, \\ \beta_1 &= \mathbf{\Sigma}_{\mathbf{zz}}^{-1} \sigma_{Y\mathbf{z}}. \end{aligned}$$

Proof: For arbitrary coefficients $(b_0, \mathbf{b}_1)$ we have the decomposition

$$\begin{aligned} & E\left[\left\{Y - b_0 - \mathbf{b}_1^T \mathbf{z}\right\}^2\right] \\ = \; & E\left[\left\{ \begin{array}{c} \left(Y - \beta_0 - \beta_1^T \mathbf{z}\right) \\ -\left((b_0 - \beta_0) + (\mathbf{b}_1 - \beta_1)^T \mathbf{z}\right) \end{array} \right\}^2\right] \\ = \; & E\left[\left(Y - \beta_0 - \beta_1^T \mathbf{z}\right)^2\right] \\ & + E\left[\left((b_0 - \beta_0) + (\mathbf{b}_1 - \beta_1)^T \mathbf{z}\right)^2\right], \end{aligned}$$

where the final equality is because $Y - \beta_0 - \beta_1^T \mathbf{z}$ has a mean of 0 and is uncorrelated with $\mathbf{z}$. □

- A related problem: find the linear combination $\mathbf{b}^T \mathbf{z}$ which is most highly correlated with Y. This correlation is (with $\mathbf{a} = \boldsymbol{\Sigma}_{\mathbf{zz}}^{1/2} \mathbf{b}$, then using the Cauchy-Schwarz Inequality)

$$\frac{\sigma_{Y\mathbf{z}}^T \mathbf{b}}{\sqrt{\sigma_{YY}} \sqrt{\mathbf{b}^T \boldsymbol{\Sigma}_{\mathbf{zz}} \mathbf{b}}} = \frac{\sigma_{Y\mathbf{z}}^T \boldsymbol{\Sigma}_{\mathbf{zz}}^{-1/2} \mathbf{a}}{\sqrt{\sigma_{YY}} \left\|\mathbf{a}\right\|} \le \frac{\left\|\boldsymbol{\Sigma}_{\mathbf{zz}}^{-1/2} \sigma_{Y\mathbf{z}}\right\|}{\sqrt{\sigma_{YY}}} \overset{def}{=} R,$$

with equality iff $\mathbf{a} \propto \mathbf{\Sigma}_{\mathbf{zz}}^{-1/2}\sigma_{Y\mathbf{z}}$, i.e. $\mathbf{b} \propto \beta_1$ – the best linear function is again the conditional mean. The maximum correlation R is called the *multiple correlation coefficient* for the population, and

$$R^2 = \frac{\sigma_{Y\mathbf{z}}^T \mathbf{\Sigma}_{\mathbf{zz}}^{-1} \sigma_{Y\mathbf{z}}}{\sigma_{YY}},$$

is called the *coefficient of determination*. Note that, *under normality*,

$$1 - R^2 = \frac{\sigma_{YY} - \sigma_{Y\mathbf{z}}^T \mathbf{\Sigma}_{\mathbf{zz}}^{-1} \sigma_{Y\mathbf{z}}}{\sigma_{YY}}$$

is the ratio of the conditional variance (given $\mathbf{z}$) to the unconditional variance of Y, so that $\sigma_{YY}R^2$ is the amount by which var$[Y]$ is reduced through the regression on $\mathbf{z}$.

- **Partial correlation**. Suppose that Y_1, Y_2 are each regressed on $\mathbf{z}$, resulting in the errors

$$\begin{aligned} Y_1 - \mu_{Y_1} - \sigma_{Y_1\mathbf{z}}^T \mathbf{\Sigma}_{\mathbf{zz}}^{-1} \left(\mathbf{z} - \mu_{\mathbf{z}}\right), \\ Y_2 - \mu_{Y_2} - \sigma_{Y_2\mathbf{z}}^T \mathbf{\Sigma}_{\mathbf{zz}}^{-1} \left(\mathbf{z} - \mu_{\mathbf{z}}\right). \end{aligned}$$

The covariance matrix of these two errors is

$$\mathbf{\Sigma_Y} - \mathbf{\Sigma_{Yz}}\mathbf{\Sigma_{zz}^{-1}}\mathbf{\Sigma_{zY}} \overset{def}{=} \begin{pmatrix} \sigma_{Y_1 Y_1 \cdot \mathbf{z}} & \sigma_{Y_1 Y_2 \cdot \mathbf{z}} \\ \sigma_{Y_1 Y_2 \cdot \mathbf{z}} & \sigma_{Y_2 Y_2 \cdot \mathbf{z}} \end{pmatrix}, \tag{13.1}$$

and the resulting correlation

$$\rho_{Y_1 Y_2 \cdot \mathbf{z}} = \frac{\sigma_{Y_1 Y_2 \cdot \mathbf{z}}}{\sqrt{\sigma_{Y_1 Y_1 \cdot \mathbf{z}}}\sqrt{\sigma_{Y_2 Y_2 \cdot \mathbf{z}}}}$$

is the *partial correlation coefficient* between Y_1 and Y_2, after eliminating the effect of $\mathbf{z}$.

- The mles of these quantities are obtained by replacing the parameters by their mles; the invariance of the mle is used here. If the joint distributions are normal, then (the estimate of) R^2 is the R^2 returned by every regression package. The matrix (13.1) is the conditional covariance cov$[\mathbf{Y}|\mathbf{z}]$, and the mle of $\rho_{Y_1 Y_2 \cdot \mathbf{z}}$ becomes the sample correlation between two sets of residuals – ‘after adjusting for the effect of $\mathbf{z}$’.

Part III

ANALYSIS OF COVARIANCE STRUCTURES

14. Principal components – theory

- **Population principal components**. For a population of random vectors $\mathbf{x} \sim (\mu, \Sigma)$ (not necessarily normal) we aim to find linear combinations $\mathbf{a}^T\mathbf{x}$ explaining 'most' of the variation. (What properties distinguish the members of the population from each other? Example: $\mathbf{x}$ represents measures on various companies from two sectors – oil and financial.) We impose the restriction $\|\mathbf{a}\| = 1$ (why?), and require that these linear combinations be uncorrelated. Let

$$\boldsymbol{\Sigma} = \mathsf{\Gamma}\mathsf{\Lambda}\mathsf{\Gamma}^T$$

be the spectral decomposition, with

$$\mathsf{\Lambda} = diag\left(\lambda_1 \geq \cdots \geq \lambda_p\right),$$

the diagonal matrix of ordered eigenvalues and

$$\mathsf{\Gamma} = \left(\gamma_1 \vdots \cdots \vdots \gamma_p\right),$$

the orthogonal matrix of corresponding eigenvectors. Then (with $\mathbf{a} = \mathsf{\Gamma}\mathbf{b}$; $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$)

$$\text{var}\left[\mathbf{a}^T\mathbf{x}\right] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} = \mathbf{b}^T\mathsf{\Lambda}\mathbf{b}.$$

This variance is maximized by (with $\mathbf{e}_i =$ ..., as usual)

$$\begin{aligned} \mathbf{b}_1 &= \mathbf{e}_1, \\ \mathbf{a}_1 &= \boldsymbol{\gamma}_1. \end{aligned}$$

(How?). We call $\boldsymbol{\gamma}_1^T \mathbf{x}$ the *first (population) principal component*. A linear combination $\mathbf{a}_2^T \mathbf{x}$ uncorrelated with $\mathbf{a}_1^T \mathbf{x}$ necessarily has $\mathbf{b}_2 \perp \mathbf{e}_1$ (how?). Then the first element of $\mathbf{b}_2$ must vanish, and so

$$\mathsf{var}\left[\mathbf{a}_2^T \mathbf{x}\right] = \mathbf{b}_2^T \Lambda \mathbf{b}_2 = \sum_{i=2}^{p} \lambda_i b_{2,i}^2;$$

this is maximized by

$$\begin{aligned} \mathbf{b}_2 &= \mathbf{e}_2, \\ \mathbf{a}_2 &= \boldsymbol{\gamma}_2. \end{aligned}$$

Continuing, we obtain the principal components

$$\left\{ Y_j = \boldsymbol{\gamma}_j^T \mathbf{x};\ j = 1, ..., p \right\}.$$

These are uncorrelated, with variances $\left\{\lambda_j\right\}$:

$$\begin{aligned}\text{cov}\left[Y_j, Y_k\right] &= \boldsymbol{\gamma}_j^T \boldsymbol{\Sigma} \boldsymbol{\gamma}_k \\ &= \mathbf{e}_j^T \Lambda \mathbf{e}_k \\ &= \Lambda_{jk} = \begin{cases} \lambda_j, & j = k, \\ 0, & j \neq k. \end{cases}\end{aligned}$$

- The pc's $\mathbf{y} = (Y_1, ..., Y_p)^T$ contain the same information as $\mathbf{x}$:

  $$\mathbf{y} = \Gamma^T \mathbf{x} \text{ and } \mathbf{x} = \Gamma \mathbf{y}, \qquad (14.1)$$

  but the intention is that the first few elements of $\mathbf{y}$ summarize this information adequately.

- Correlations between X_i and Y_j: denoting by $\mathbf{r}_i^T = \left(\gamma_{i1}, ..., \gamma_{ip}\right)$ the i^{th} row of Γ, and using the second equality in (14.1), gives

  $$\begin{aligned}\text{cov}\left[X_i, Y_j\right] &= \text{cov}\left[\mathbf{r}_i^T \mathbf{y}, \mathbf{y}^T \mathbf{e}_j\right] \\ &= \mathbf{r}_i^T \Lambda \mathbf{e}_j \\ &= \lambda_j \mathbf{r}_i^T \mathbf{e}_j \\ &= \lambda_j \gamma_{ij},\end{aligned}$$

(where $\gamma_{ij} = \Gamma_{ij} = i^{th}$ element of $\boldsymbol{\gamma}_j$) and so

$$\begin{aligned} \text{corr}\left[X_i, Y_j\right] &= \frac{\lambda_j \gamma_{ij}}{\sqrt{\sigma_{ii}\lambda_j}} \\ &= \gamma_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}} \\ &= \rho_{X_i,Y_j}, \text{ say.} \end{aligned}$$

But

$$\begin{aligned} \sigma_{ii} &= \text{var}\left[\mathbf{r}_i^T \mathbf{y}\right] \\ &= \mathbf{r}_i^T \Lambda \mathbf{r}_i \\ &= \sum_j \lambda_j \gamma_{ij}^2 \\ &= \sigma_{ii} \sum_j \rho_{X_i,Y_j}^2 . \end{aligned}$$

Thus $\sum_j \rho_{X_i,Y_j}^2 = 1$ and we interpret ρ_{X_i,Y_j}^2 as the proportion of the variance of X_i 'explained by (the correlation with) Y_j'.

- If $\mathbf{x} \sim N(\mu, \boldsymbol{\Sigma})$ then the pc's are independently and normally distributed: $Y_j \sim N\left(\boldsymbol{\gamma}_j^T \mu, \lambda_j\right)$.

- The proportion of the total variation ($= tr\mathbf{\Sigma}$, not merely the variation of one X_i) explained by the first q pc's is
$$\frac{\lambda_1 + \cdots + \lambda_q}{tr\mathbf{\Sigma}} = \frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_q + \cdots + \lambda_p}.$$
This is a guide to the number of pc's to be studied.

- The pc's are not scale invariant. (What does this mean and why is it undesirable?) As well, especially when the variances σ_{jj} are themselves highly varied, or when the variables are expressed in different units, it can be easier to interpret the pc's obtained from the standardized variables
$$\mathbf{z} = \mathbf{V}^{-1/2}(\mathbf{x} - \mu),$$
where $\mathbf{V} = diag(\sigma_{11}, \cdots, \sigma_{pp})$. Note that the elements of $\mathbf{z}$ are dimensionless. Since
$$\text{cov}[\mathbf{z}] = \mathbf{V}^{-1/2}\mathbf{\Sigma}\mathbf{V}^{-1/2} \overset{def}{=} \mathbf{P},$$
(with elements $\rho_{ij} =$corr$\left[X_i, X_j\right]$) the corresponding pc's are $\left\{\tilde{Y}_j = \tilde{\gamma}_j^T\mathbf{z};\ j = 1, ..., p\right\}$, where the

$\tilde{\gamma}_j$ are the eigenvectors of $\mathbf{P}$. The variances of the $\tilde{Y}_j$ are the eigenvalues of $\mathbf{P}$. Note that $tr\mathbf{P} = p$.

- The sample principal components $\left\{\hat{Y}_j\right\}$ are obtained by carrying out the analysis, described above, on the sample covariance matrix. They are generally computed after standardizing, and then computing the eigenvalue-eigenvector decomposition $\left(\hat{\Lambda}, \hat{\Gamma}\right)$ of the sample correlation matrix $\mathbf{R}$. Thus, with

$$\mathbf{X}_0 = \begin{pmatrix} \vdots \\ \frac{x_{i1}-\bar{x}_1}{\sqrt{s_{11}}}, \cdots, \frac{x_{ip}-\bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots \end{pmatrix},$$

we have

$$\mathbf{R} = \frac{1}{n-1}\mathbf{X}_0^T\mathbf{X}_0$$

and then (recall that $\mathbf{y}^T = \mathbf{x}^T\Gamma$ in the population) the sample pc's are the columns $\hat{\mathbf{y}}^{(1)}$, ..., $\hat{\mathbf{y}}^{(p)}$ of

$$\mathbf{Y}_{n\times p} = \mathbf{X}_0\hat{\Gamma}.$$

The elements of these columns are the 'scores' – one for each of the n standardized observations.

Each of these pc's has an average of 0 (why?), and their sample covariance is

$$\begin{aligned}\frac{1}{n-1}\mathbf{Y}^T\mathbf{Y} &= \frac{1}{n-1}\hat{\Gamma}^T\mathbf{X}_0^T\mathbf{X}_0\hat{\Gamma}\\ &= \hat{\Gamma}^T\mathbf{R}\hat{\Gamma}\\ &= \hat{\Lambda},\end{aligned}$$

i.e. the sample pc's have sample correlations of 0 and the sample variance of the j^{th} of them is $\hat{\lambda}_j$.

- Some large sample approximations to the distributions are available. In particular, if $\lambda_1 > \cdots > \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}$ (i.e. all are distinct) and $\left\{\hat{\lambda}_j\right\}$ the eigenvalues of $\mathbf{S}$, then

  $$\sqrt{n}\left(\hat{\lambda}-\lambda\right)\xrightarrow{L} N\left(0, 2\Lambda^2\right),$$

  i.e. the $\hat{\lambda}_j$ are asymptotically normal and independent with asymptotic variances $2\lambda_j^2/n$. Then by the delta method, the r.v.s $\left\{\log\hat{\lambda}_j - \log\lambda_j\right\}$ are asymptotically normal and independent, with equal asymptotic variances of $2/n$. This result can be used to obtain confidence intervals.

15. Principal components – examples

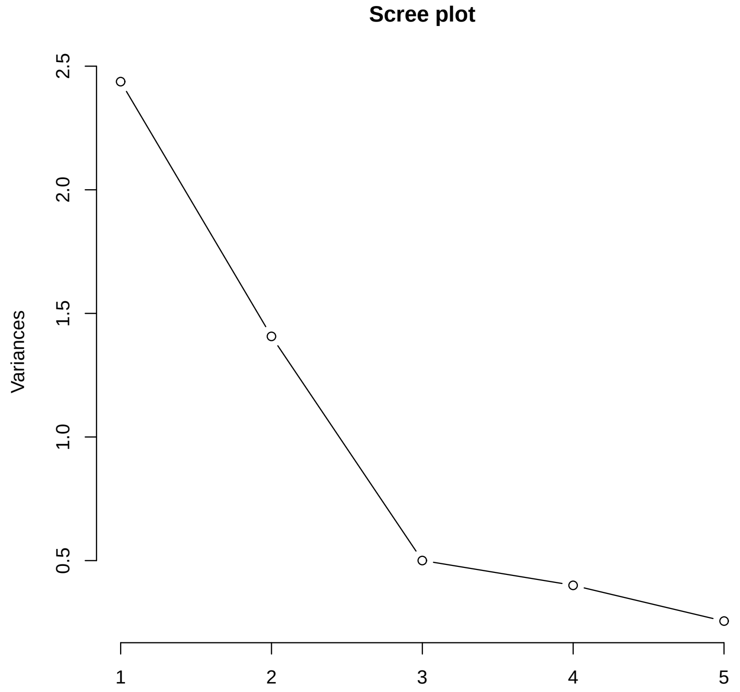
- **Example 1**: e.g. 8.5; data in Table 8.4.

```
out = prcomp(X, scale = T)
   yields
Standard deviations:
[1] 1.561 1.186 0.707 0.632 0.505
 # Square roots of the eigenvalues of R

Rotation:
          PC1    PC2    PC3    PC4    PC5
JPM    -0.469  0.368 -0.604  0.363  0.384
Citi   -0.532  0.236 -0.136 -0.629 -0.496
WellsF -0.465  0.315  0.772  0.289  0.071
Shell  -0.387 -0.585  0.093 -0.381  0.595
Exxon  -0.361 -0.606 -0.109  0.493 -0.498
  # Columns are eigenvectors of R,
  # (= the coefficients of the sample pc’s.)

cumsum(out$sdev^2)/5
[1] 0.487 0.769 0.869 0.949 1.000
```

```
plot(out)
```



- Interpretation of first pc is the normalized sum – general market activity. Second represents a contrast between financial firms and oil firms:

$$\begin{aligned} \mathbf{a}_1 &\approx (1,1,1,1,1)^T / \|\cdots\|, \\ \mathbf{a}_2 &\approx (1/3,1/3,1/3,-1/2,-1/2)^T / \|\cdots\|. \end{aligned}$$

- $\hat{\lambda}_1 = 2.437$; CI on $\log \lambda_1$ is $\log \hat{\lambda}_1 \pm z_{\alpha/2}\sqrt{2/n}$ [IF the asymptotic theory using $\mathbf{S}$ applies to $\mathbf{R}$]; exponentiate to get ci on λ_1:

```
[1] 1.854779 3.202700
```

Similarly for λ_2:

```
[1] 1.070745 1.848886
```

- **Note**: Even when `scale = F`, the data will be centred by R – the column averages will be subtracted; thus the scores still have an average of zero.

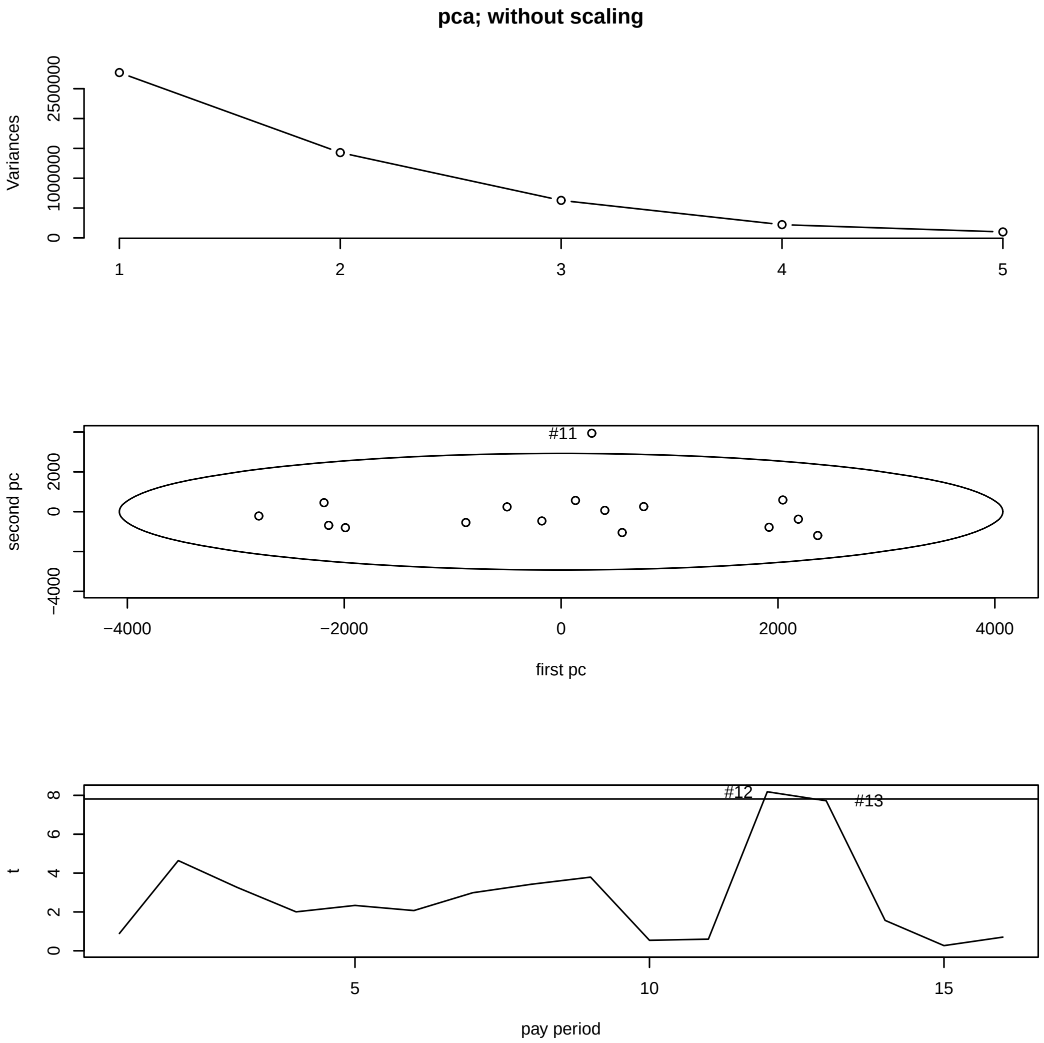
- **Example 2**: Data (Table 5.8) on police overtime expenses; $p = 5$ (index represents time):

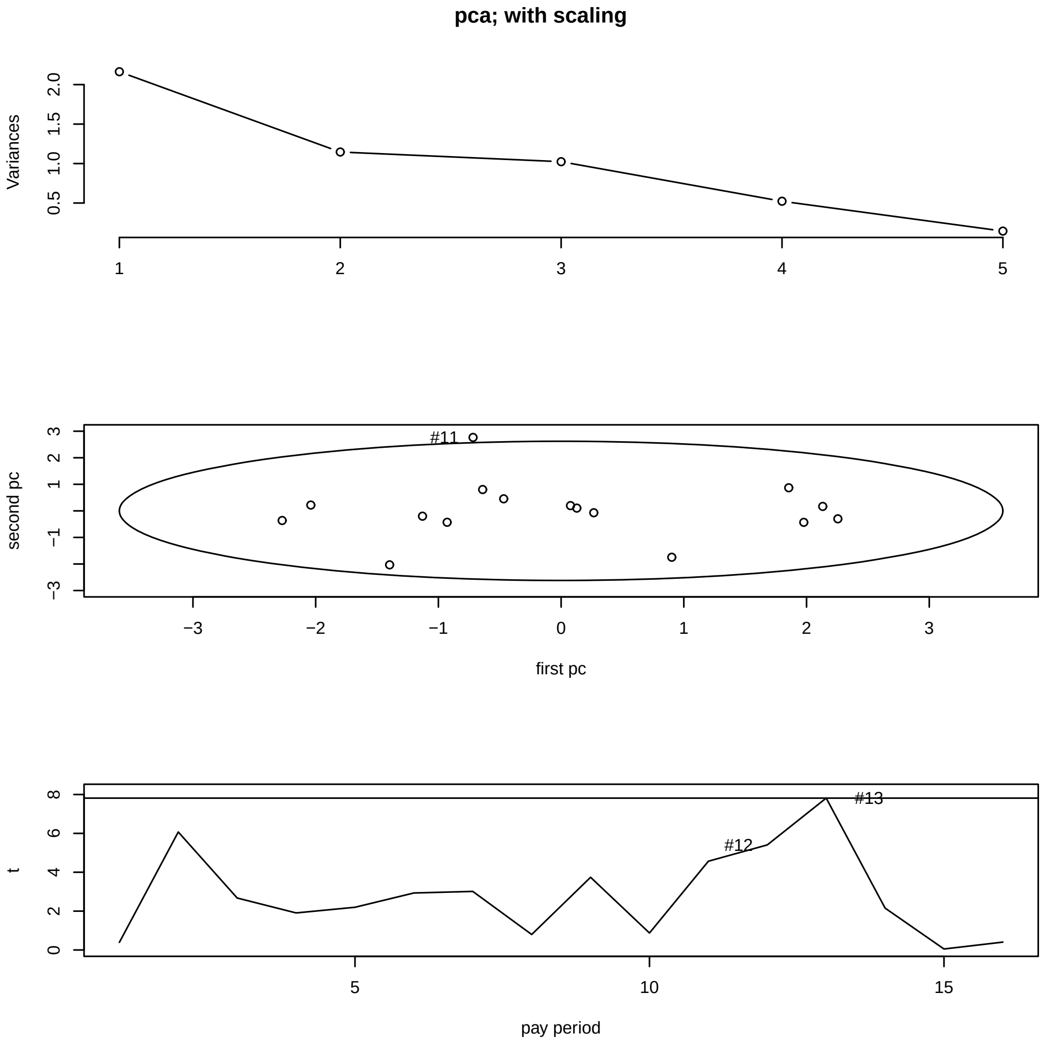
$$\mathbf{x} = \begin{pmatrix} \text{legal appearances} \\ \text{extraordinary event} \\ \text{holdover} \\ \text{COA} \\ \text{meetings} \end{pmatrix}.$$

- To visually check for outliers, plot the first two pc's and the 95% ellipsoid

$$\frac{y_1^2}{\hat{\lambda}_1} + \frac{y_2^2}{\hat{\lambda}_2} \leq \chi_2^2\,(.05)\,.$$

  Why are we using the χ_2^2 distribution (as an approximation)? Code and output on web site. Observation $\#11$ seems to be an outlier; indeed its X_2-value is very large (and the second pc is dominated by X_2; it explains almost all of the variation of this variable).


pca; without scaling
Variances
0
1000000
2500000
1
2
3
4
5
second pc
#11
2000
0
−4000
−4000
−2000
0
2000
4000
first pc
t
8
6
4
2
0
#12
#13
5
10
15
pay period


pca; with scaling
Variances
0.5
1.0
1.5
2.0
1
2
3
4
5
second pc
−3
−1
1
2
3
#11
−3
−2
−1
0
1
2
3
first pc
t
0
2
4
6
8
#12
#13
5
10
15
pay period

- To check for ‘out of control’ observations due to more minor causes, we note that

  $$\sum_{j=3}^{p} \frac{y_j^2}{\hat{\lambda}_j} \overset{d}{\approx} \chi^2_{p-2}.$$

  Plotting these, and the upper limit of $\chi^2_3(.05)$, indicates the process was out of control at observations 12, 13 – right after the extraordinary overtime event.

- This uses the unscaled covariance $\mathbf{S}$; using $\mathbf{R}$ gives similar but less significant results. (Does the asymptotic theory using $\mathbf{R}$ need to be modified?)

16. Factor analysis – model and estimation

- Motivation: Suppose that variables are grouped, with those in the same group being highly correlated with each other and largely uncorrelated with those in other groups. We might try to explain the within-group correlations in terms of a small number of unobserved, random *factors*. Example: highly correlated scores in academic subjects, and highly correlated performances in athletic activities, might be explained by two factors – intelligence and athletic ability. The choices made here are often controversial (might social or cultural factors account just as well for the groupings?).

- Orthogonal factor model:

$$\mathbf{x}_{p\times 1} = \mu + \mathbf{L}_{p\times m}\mathbf{f}_{m\times 1} + \varepsilon. \tag{16.1}$$

The elements $F_1, ..., F_m$ of $\mathbf{f}$ are r.v.s called *common factors*; $\mathbf{L}$ is a non-random matrix of *factor* *loadings* with rank $m \leq p$. The mean-zero random errors (‘specific factors’) ε_i are uncorrelated with each other and with the F_j; we assume as well that

$$\begin{aligned} E[\mathbf{f}] &= \mathbf{0},\ \text{cov}[\mathbf{f}] = \mathbf{I}_m, \\ \text{cov}[\varepsilon] &= \mathbf{\Psi} = diag(\psi_1, ..., \psi_p). \end{aligned}$$

That cov$[\mathbf{f}] = \mathbf{I}_m$ motivates the name ‘orthogonal factors’. The diagonal elements of $\mathbf{\Psi}$ are termed ‘uniquenesses’ or ‘specific variances’.

- The structure implies that

  $$\text{cov}[\mathbf{x}] = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi} \text{ and } \text{cov}[\mathbf{x}, \mathbf{f}] = \mathbf{L}.$$

  Then cov$\left[X_i, F_j\right] = l_{ij}$ – the “loading of the i^{th} variable on the j^{th} factor” – and

  $$\text{var}[X_i] = \sigma_{ii} = h_i^2 + \psi_i,$$

  where the “i^{th} communality” – the part of the variance arising from the factor loadings – is

  $$h_i^2 \overset{def}{=} \left(\mathbf{L}\mathbf{L}^T\right)_{ii}.$$

- A simple example is $p = 3$, one factor ($m = 1$), $\mathbf{L} = \mathbf{1}_p$,
$$\mathbf{\Sigma} = \text{cov}\left[\mathbf{x}\right] = \left(\begin{array}{ccc} 1+\psi_1 & 1 & 1 \\ 1 & 1+\psi_2 & 1 \\ 1 & 1 & 1+\psi_3 \end{array} \right).$$
Of course not every covariance matrix $\mathbf{\Sigma}$ can be written as $\mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$ with $\mathbf{\Psi}$ a diagonal variance matrix (if $m < p$ is given).

- The model is scale invariant in that if $\mathbf{x}$ follows model (16.1) then so does $\mathbf{y} = \mathbf{C}\mathbf{x}$ with $\mathbf{C}$ diagonal (required so that cov$[\mathbf{C}\varepsilon] = \mathbf{C}\mathbf{\Psi}\mathbf{C}^T$ is diagonal).

- There is an identifiability problem: given an $m \times m$ orthogonal matrix Γ, we can replace $\mathbf{L}$ by $\mathbf{L}\Gamma$ and $\mathbf{f}$ by $\Gamma^T\mathbf{f}$ without altering the structure of $\mathbf{\Sigma}$. One typically ‘rotates’ the matrix $\mathbf{L}$, choosing a particular version $\mathbf{L}\Gamma$ according to an ‘ease of interpretation’ criterion.

- **Principal component solution**. One (approximate) factorization of the covariance matrix starts by representing it as

$$\mathbf{\Sigma} = \Gamma\Lambda\Gamma^T = \sum_{j=1}^{p} \lambda_j \gamma_j \gamma_j^T,$$

where $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues and $\left\{\gamma_j\right\}$ the eigenvectors, and approximating this as

$$\begin{aligned} \mathbf{\Sigma} &\approx \sum_{j=1}^{m} \lambda_j \gamma_j \gamma_j^T = \mathbf{L}\mathbf{L}^T, \text{ for} \\ \mathbf{L} &= \left(\sqrt{\lambda_1}\gamma_1 \vdots \cdots \vdots \sqrt{\lambda_m}\gamma_m\right). \end{aligned}$$

Then the approximation is improved by adding in $\mathbf{\Psi}$, with diagonal elements $\psi_i = \sigma_{ii} - \left(\mathbf{L}\mathbf{L}^T\right)_{ii}$, i.e.

$$\mathbf{\Psi} = diag\left(\mathbf{\Sigma} - \mathbf{L}\mathbf{L}^T\right).$$

Given p-dimensional data $\mathbf{x}_1, ..., \mathbf{x}_n \sim (\mu, \mathbf{\Sigma})$ this is applied to the covariance matrix $\mathbf{S}_{p\times p}$ (or the correlation matrix $\mathbf{R}$), yielding estimates $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ and a residual matrix

$$\mathbf{S} - \left(\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}\right).$$

Using the 'Frobenius' norm $\|\mathbf{A}\|^2 = tr\mathbf{A}^T\mathbf{A} = \sum_{i,j} a_{ij}^2$ $(= \|vec\mathbf{A}\|^2)$, we have (assigned)

$$\left\|\mathbf{S} - \left(\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}\right)\right\|^2 \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2; \tag{16.2}$$

a guide to the choice of m.

- The contribution of the j^{th} common factor towards the variance s_{ii} is $\hat{l}_{ij}^2$ (how?), and so its (proportional) contribution towards the total variance is

$$\frac{\sum_{i=1}^p \hat{l}_{ij}^2}{tr\mathbf{S}} = \frac{\left\|\sqrt{\hat{\lambda}_j}\hat{\gamma}_j\right\|^2}{tr\mathbf{S}} = \frac{\hat{\lambda}_j}{tr\mathbf{S}}, \tag{16.3}$$

a better guide to the choice of m.

- Example – e.g. 8.5 again, discussed as e.g. 9.4 – stock price data. See web site: `princomp(X, cor=TRUE)` starts it off.

- **Maximum likelihood solution**. Here one assumes normality, and maximizes the likelihood over μ (obtaining the mle $\bar{\mathbf{x}}$) and $\boldsymbol{\Sigma}$, subject to $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi}$. Because of the lack of identifiability a further constraint must be imposed; most common is that $\mathbf{L}^T\boldsymbol{\Psi}^{-1}\mathbf{L} = \boldsymbol{\Delta} = diag\,(\delta_1 \geq \cdots \geq \delta_m)$. Under this constraint $\mathbf{L}$ is uniquely defined (up to multiplication of its columns by ± 1) if as well the δ's are distinct (true w.p. 1 when applied to the data).

  - **Reason**: First note that if $\mathbf{L}$ is not uniquely defined, then there are $\mathbf{L}_{p\times m}$ and $\mathbf{K}_{p\times m}$ with $\mathbf{L}\mathbf{L}^T = \mathbf{K}\mathbf{K}^T$. Then necessarily

    $$\mathbf{K} = \mathbf{L}\boldsymbol{\Gamma} \text{ for some orthogonal } \boldsymbol{\Gamma}.$$

    Why? By the singular value decomposition (proof to follow) we can write $\mathbf{L} = \mathbf{G}\mathbf{D}\boldsymbol{\Gamma}_L$, where $\mathbf{G}_{p\times p}$ is the orthogonal matrix of eigenvectors of $\mathbf{L}\mathbf{L}^T$ (hence of $\mathbf{K}\mathbf{K}^T$), $\mathbf{D}_{p\times m}$ has the 'singular values of $\mathbf{L}$' – the roots of the nonzero eigenvalues of $\mathbf{L}\mathbf{L}^T$ (hence of $\mathbf{K}\mathbf{K}^T$)

– on its diagonal and zeros elsewhere, and Γ_L is the orthogonal matrix of eigenvectors of $\mathbf{L}^T\mathbf{L}$. Thus we can write $\mathbf{K} = \mathbf{GD}\Gamma_K = \mathbf{L}\Gamma$, where $\Gamma = \Gamma_L^T \Gamma_K$ is orthogonal.

Now, if as well $\mathbf{L}^T\Psi^{-1}\mathbf{L} = \Delta$ (*) and $\mathbf{K}^T\Psi^{-1}\mathbf{K} = \tilde{\Delta}$, with Δ and $\tilde{\Delta}$ diagonal, then $\tilde{\Delta} = \Gamma^T\Delta\Gamma$. Thus Δ and $\tilde{\Delta}$ have the same eigenvalues – i.e. the same diagonal elements. If these diagonal elements are ordered as described then $\Delta = \tilde{\Delta}$ and so $\Gamma\Delta = \Delta\Gamma$. Comparing the $(i,j)^{th}$ elements gives $\gamma_{ij} = 0$ if $i \neq j$, hence Γ is diagonal with diagonal elements ± 1. [One can replace Ψ in (*) by any invertible matrix; another common choice is $diag\,(\Sigma)$.]

17. Factor analysis – rotation and scores

- Example of ML fit of factor analysis model on web site. The R function `factanal()` computes the mle on the basis of the sample correlation matrix $\mathbf{R}$. In the pc solution all variables have positive loadings on F_1 – a 'market factor' – with the banking/oil stocks having positive/negative loadings – 'industry factor'? The (unrotated) ml solution perhaps has an interpretation as a 'banking factor' and 'oil industry factor'.

- The first term in (16.3) – proportional SS of the loadings on the j^{th} factor – still represents the proportional contribution of the j^{th} common factor towards the total variance; estimated by replacing the terms by their mles (= `Proportion Var` in the R output).

- There is a likelihood ratio test for m together with a Bartlett's correction; R includes in the output the p-value of this test, for the value of m used. If m is too large there are more parameters in the restricted FA model than in the unrestricted model and the mle cannot be computed. This occurs if the decrease in the number of parameters $\left(= \left[(p-m)^2 - (p+m)\right]/2\right)$ is ≤ 0; in fact the LR χ^2 is on $\left[(p-m)^2 - (p+m)\right]/2$ df.

  – Reason: there are $p\left(p+1\right)/2$ parameters in the unrestricted model; in the FA model there are the p uniquenesses and the pm parameters in $\mathbf{L}$, subject to $\left(m-1\right)m/2$ constraints – the upper triangle of $\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L}$ must vanish.

- The FA model is invariant under orthogonal transformations $\mathbf{L} \rightarrow \mathbf{L}\Gamma = \tilde{L}$, with $\Gamma_{m\times m}$ orthogonal. One generally attempts to choose a transformation to make the factors more interpretable, e.g. so that the loadings on each particular factor split into interpretable groups – those with 'large' loadings, and the others (for each $j = 1,..,m$, $\tilde{l}_{ij}^2$ should be highly varied).

- **Varimax criterion**. Note that

  $$h_i^2 = \sum_j l_{ij}^2 = \left(\mathbf{L}\mathbf{L}^T\right)_{ii} = \left(\tilde{L}\tilde{L}^T\right)_{ii} = \sum_j \tilde{l}_{ij}^2;$$

  define

  $$v_{ij} = \tilde{l}_{ij}^2 / h_i^2,\ (\sum_j v_{ij} = 1).$$

  The varimax criterion seeks a transformation which maximizes their variances $\left\{s_j^2\right\}$:

  $$\sum_{j=1}^{m} \left\{ \frac{1}{p-1} \sum_{i=1}^{p} \left(v_{ij} - \bar{v}_{\cdot j}\right)^2 \right\} = \sum_j s_j^2.$$

- Example – stock-price data; on web site. Interpretation – factor 1 represents economic forces causing bank stocks to move together; factor 2 such forces affecting oil stocks – ??? – highly subjective.

  - `factanal()` will by default rotate the maximum likelihood solution; to rotate the principal component solution use `varimax(L)`, where

$\mathbf{L}$ is the pc loadings matrix. Can also use `varimax(loadings(fit))`, where `fit` is the ml solution (but not the pc solution – try it to see why).

- **Estimation of $\mathbf{f}$ (factor scores)**. We are seeking estimates of the values attained by these, random, factors. The methods to be described treat the estimates of $\mathbf{L}$ and $\mathbf{\Psi}$ as known; the estimates are then 'plugged-in'.

  – **Weighted least squares method** ('Bartlett' scores on R, when using `factanal()`). If $\mathbf{L}$, $\mathbf{\Psi}$ and μ are known, then

$$\mathbf{x} - \mu = \mathbf{L}\mathbf{f} + \varepsilon,$$

with $\varepsilon \sim (\mathbf{0}, \mathbf{\Psi})$. The wls estimate of $\mathbf{f}$ is

$$\begin{aligned} \hat{\mathbf{f}}_{m\times 1} &= \text{arg min}\,(\mathbf{x} - \mu - \mathbf{L}\mathbf{f})^T\,\mathbf{\Psi}^{-1}\,(\mathbf{x} - \mu - \mathbf{L}\mathbf{f}) \\ &= \left(\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{\Psi}^{-1}\,(\mathbf{x} - \mu)\,. \end{aligned}$$

This is applied as

$$\hat{\mathbf{f}}_i = \left(\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\,(\mathbf{x}_i - \bar{\mathbf{x}})\,, i = 1, ..., n.$$

If the loadings are rotated – $\mathbf{L} \rightarrow \mathbf{L}\Gamma$ – then the scores are correspondingly transformed: $\hat{\mathbf{f}} \rightarrow \Gamma^T \hat{\mathbf{f}}$ (so that $\mathbf{L}\mathbf{f}$ remains fixed).

∗ The 'factor score coefficients' are the m rows of $\left(\hat{\mathbf{L}}^T \hat{\mathbf{\Psi}}^{-1} \hat{\mathbf{L}}\right)^{-1} \hat{\mathbf{L}}^T \hat{\mathbf{\Psi}}^{-1}$ – one for each factor; the $\hat{\mathbf{f}}_i$ are the n columns of $\left[\left(\hat{\mathbf{L}}^T \hat{\mathbf{\Psi}}^{-1} \hat{\mathbf{L}}\right)^{-1} \hat{\mathbf{L}}^T \hat{\mathbf{\Psi}}^{-1}\right] \mathbf{X}^T_{centred}$: $m \times n$.

∗ When the pc solution is used, it is common to take $\hat{\mathbf{\Psi}} \propto \mathbf{I}_p$; then the $\hat{\mathbf{f}}_i$ are the first m sample pcs, normalized to have sample variance $= 1$:

$$\hat{\mathbf{f}}_i = \begin{pmatrix} \frac{\hat{\gamma}_1^T (\mathbf{x}_i - \bar{\mathbf{x}})}{\sqrt{\hat{\lambda}_1}} \\ \vdots \\ \frac{\hat{\gamma}_m^T (\mathbf{x}_i - \bar{\mathbf{x}})}{\sqrt{\hat{\lambda}_m}} \end{pmatrix}.$$

∗ Example – stock price data – rotated pc solution similar to rotated ml solution with the signs of factor 2 loadings reversed.

- **Regression method** ('regression' scores on R). If the distributions are Normal, then $\mathbf{x}-\mu$ and $\mathbf{f}$ are jointly normal:

  $$\begin{pmatrix} \mathbf{x}-\mu \\ \mathbf{f} \end{pmatrix} \sim N_{p+m}\left(\mathbf{0}, \begin{pmatrix} \mathbf{\Sigma}=\mathbf{L}\mathbf{L}^T+\mathbf{\Psi} & \mathbf{L} \\ \mathbf{L}^T & \mathbf{I}_m \end{pmatrix}\right).$$

  The conditional distribution of $\mathbf{f}$, given $\mathbf{x}$, is

  $$N_m\left(\mathbf{L}^T\mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mu\right), \mathbf{I}_m - \mathbf{L}^T\mathbf{\Sigma}^{-1}\mathbf{L}\right),$$

  and $\mathbf{f}$ is estimated by the conditional mean:

  $$\hat{\mathbf{f}}_i = \hat{\mathbf{L}}^T\hat{\mathbf{\Sigma}}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right).$$

  One approach here – aimed at robustness against an incorrect choice of m – uses $\hat{\mathbf{\Sigma}}=\mathbf{S}$, the unconstrained sample covariance matrix:

  $$\hat{\mathbf{f}}_i = \hat{\mathbf{L}}^T\mathbf{S}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right).$$

  Another proceeds by using the matrix identity

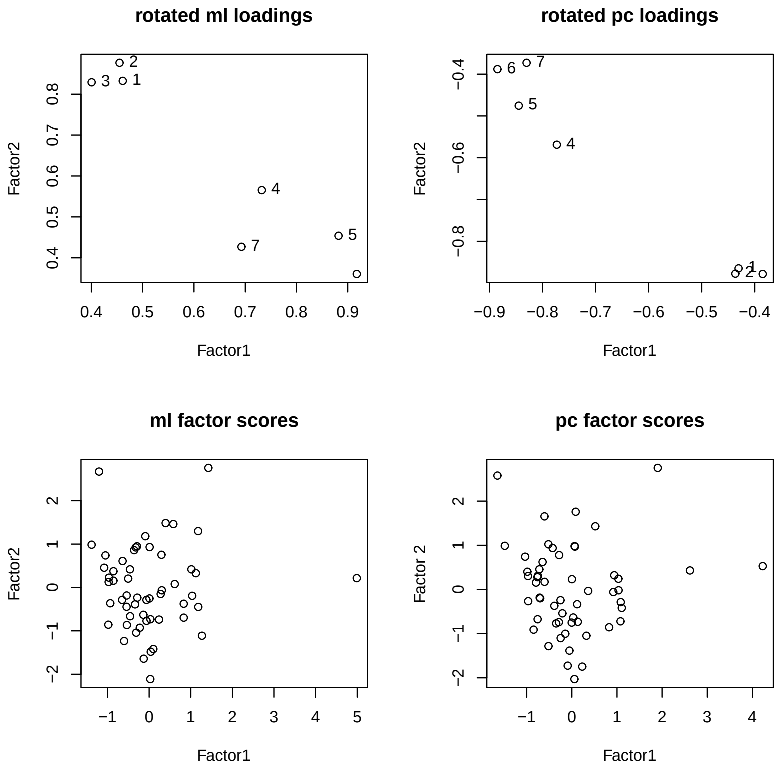
  $$\begin{aligned} \mathbf{L}^T\mathbf{\Sigma}^{-1} &= \mathbf{L}^T\left(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi}\right)^{-1} \\ &= \left(\mathbf{I}_m+\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{\Psi}^{-1}, \quad (17.1) \end{aligned}$$

  (assigned) obtaining

  $$\hat{\mathbf{f}}_i = \left(\mathbf{I}_m+\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right).$$

- Example – stock price data

- Example – Olympic decathlon data. Variables 1:100m, 2:200m, 3:400m, 4:800m, 5:1500m, 6:3000m, 7:mrthn.



Top: Factor loadings – the columns of $\mathbf{L}$ – plotted against each other; $m = 2$: factors might represent running speed and endurance? Bottom: Factor scores (pc scores $\times - 1$).

18. Canonical correlations I

- **Singular Value Decomposition (SVD)**: Let $\mathbf{K}_{p\times m}$ have rank r $\left(= rk\left(\mathbf{K}\mathbf{K}^T\right) = rk\left(\mathbf{K}^T\mathbf{K}\right)\right)$. Define $\mathbf{D}_{r\times r} = diag\left(d_1, ..., d_r\right)$, where $d_1^2 \geq \cdots \geq d_r^2$ are the nonzero eigenvalues of $\mathbf{K}\mathbf{K}^T$, hence of $\mathbf{K}^T\mathbf{K}$. Then there are orthogonal matrices $\mathbf{G}_{p\times p}, \mathbf{H}_{m\times m}$ consisting of eigenvectors of $\mathbf{K}\mathbf{K}^T$ and $\mathbf{K}^T\mathbf{K}$ respectively, for which

$$\mathbf{K} = \mathbf{G}\left(\begin{array}{cc} \mathbf{D}_{r\times r} & \mathbf{0}_{r\times m-r} \\ \mathbf{0}_{p-r\times r} & \mathbf{0}_{p-r\times m-r} \end{array}\right)\mathbf{H}^T. \tag{18.1}$$

  - This was applied earlier in the case $r = m < p$, in which case the matrix, called $\mathbf{D}$ there, is now $\left(\begin{array}{c} \mathbf{D}_{m\times m} \\ \mathbf{0}_{p-m\times m} \end{array}\right)$.

  - Preliminary result: If $\mathbf{\Sigma} = \Gamma\Lambda\Gamma^T$ for Γ orthogonal and Λ diagonal then, necessarily, the λ_i are the eigenvalues and the γ_i the corresponding eigenvectors. (Why?)

- **Proof of SVD**: Represent the orthonormal eigenvectors of $\mathbf{K}\mathbf{K}^T$ as

$$\mathbf{G} = \left(\mathbf{G}_1 : p \times r \vdots \mathbf{G}_2 : p \times p - r\right),$$

arranged in such a way that

$$\mathbf{K}\mathbf{K}^T = \mathbf{G}\begin{pmatrix} \mathbf{D}^2_{r\times r} & \mathbf{0}_{r\times p-r} \\ \mathbf{0}_{p-r\times r} & \mathbf{0}_{p-r\times p-r} \end{pmatrix}\mathbf{G}^T.$$

Note that then

$$\mathbf{G}_2^T\mathbf{K}\mathbf{K}^T\mathbf{G}_2 = \mathbf{0}_{p-r\times p-r},$$

implying that $\mathbf{G}_2^T\mathbf{K} = \mathbf{0}_{p-r\times m}$ and so

$$\mathbf{G}_1\mathbf{G}_1^T\mathbf{K} = \left(\mathbf{I}_p - \mathbf{G}_2\mathbf{G}_2^T\right)\mathbf{K} = \mathbf{K}. \tag{18.2}$$

This – (18.2) – is the crucial step. Now define

$$\mathbf{H} = \left(\mathbf{H}_1 : m \times r \vdots \mathbf{H}_2 : m \times m - r\right),$$

with (forced!)

$$\mathbf{H}_1 = \mathbf{K}^T\mathbf{G}_1\mathbf{D}^{-1}.$$

Then (18.1) holds, because of (18.2).

Since

$$\begin{aligned}\mathbf{H}_1^T\mathbf{H}_1 &= \mathbf{D}^{-1}\mathbf{G}_1^T\mathbf{K}\mathbf{K}^T\mathbf{G}_1\mathbf{D}^{-1}\\ &= \mathbf{D}^{-1}\left[\mathbf{G}^T\mathbf{K}\mathbf{K}^T\mathbf{G}\right]_{11}\mathbf{D}^{-1}\\ &= \mathbf{D}^{-1}\mathbf{D}^2\mathbf{D}^{-1}\\ &= \mathbf{I}_r,\end{aligned}$$

the columns of $\mathbf{H}_1$ are orthonormal and can be augmented by those of $\mathbf{H}_2$ in such a way that $\mathbf{H}$ is orthogonal. Now

$$\begin{aligned}&\mathbf{H}\begin{pmatrix}\mathbf{D}^2_{r\times r} & \mathbf{0}_{r\times m-r}\\ \mathbf{0}_{m-r\times r} & \mathbf{0}_{m-r\times m-r}\end{pmatrix}\mathbf{H}^T\\ =\ &\mathbf{H}_1\mathbf{D}^2\mathbf{H}_1^T\\ =\ &\mathbf{K}^T\mathbf{G}_1\mathbf{G}_1^T\mathbf{K}\\ =\ &\mathbf{K}^T\mathbf{K},\end{aligned}$$

again by (18.2). Thus the columns of $\mathbf{H}$ are necessarily eigenvectors of $\mathbf{K}^T\mathbf{K}$. □

- **Population canonical correlations**: Given two random vectors $\mathbf{x}_{p\times1}, \mathbf{y}_{q\times1}$ ($p \leq q$ by convention) with covariance

$$\text{cov}\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}\right] = \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

(with $rk\left(\mathbf{\Sigma}_{11}\right) = rk\left(\mathbf{\Sigma}_{12}\right) = p$, $rk\left(\mathbf{\Sigma}_{22}\right) = q$), we wish to summarize the relationships between the two sets of variables. First find $\mathbf{a}_1, \mathbf{b}_1$ to maximize

$$\text{corr}\left[\mathbf{a}_1^T\mathbf{x}, \mathbf{b}_1^T\mathbf{y}\right] = \frac{\mathbf{a}_1^T\mathbf{\Sigma}_{12}\mathbf{b}_1}{\sqrt{\mathbf{a}_1^T\mathbf{\Sigma}_{11}\mathbf{a}_1}\sqrt{\mathbf{b}_1^T\mathbf{\Sigma}_{22}\mathbf{b}_1}}.$$

This is the *first pair of canonical variables*, or *first canonical variate pair*. The second pair is $\left(\mathbf{a}_2^T\mathbf{x}, \mathbf{b}_2^T\mathbf{y}\right)$, chosen to maximize the correlation subject to the requirement that

$$\text{corr}\left[\mathbf{a}_1^T\mathbf{x}, \mathbf{a}_2^T\mathbf{x}\right] = \text{corr}\left[\mathbf{b}_1^T\mathbf{y}, \mathbf{b}_2^T\mathbf{y}\right] = 0,$$

i.e.

$$\mathbf{a}_1^T\mathbf{\Sigma}_{11}\mathbf{a}_2 = \mathbf{b}_1^T\mathbf{\Sigma}_{22}\mathbf{b}_2 = 0.$$

Continue, ... , p pairs.

- **Solution**. First standardize: define
$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{11}^{-1/2}\mathbf{x} \\ \mathbf{\Sigma}_{22}^{-1/2}\mathbf{y} \end{pmatrix},$$
with
$$\text{cov}\left[\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}\right] = \begin{pmatrix} \mathbf{I}_p & \tilde{\mathbf{\Sigma}}_{12} \\ \tilde{\mathbf{\Sigma}}_{21} & \mathbf{I}_q \end{pmatrix},$$
where
$$\tilde{\mathbf{\Sigma}}_{12} = \mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2} = \tilde{\mathbf{\Sigma}}_{21}^T.$$
Let
$$\tilde{\mathbf{\Sigma}}_{12} = \mathbf{G}_{p\times p}\mathbf{R}_{p\times q}\mathbf{H}_{q\times q}^T$$
be the svd, with
$$\mathbf{R} = \begin{pmatrix} \rho_1^* & & 0 & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \rho_p^* & 0 & \cdots & 0 \end{pmatrix},$$
and $\rho_1^* \geq \cdots \rho_p^* > 0$. Since
$$\mathbf{R}\mathbf{R}^T = \mathbf{G}^T\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{21}\mathbf{G},$$
is diagonal, its diagonal elements ρ_i^{*2} are the eigenvalues, and the columns $\{\mathrm{g}_i\}$ of $\mathbf{G}$ are the corresponding orthonormal eigenvectors, of
$$\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{21} = \mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1/2}.$$

The ρ_i^{*2} are also the nonzero eigenvalues of $\tilde{\mathbf{\Sigma}}_{21}\tilde{\mathbf{\Sigma}}_{12}$: $q \times q$, and the columns of $\mathbf{H}$ are the corresponding orthonormal eigenvectors. Now put

$$\begin{aligned}\tilde{u}_{p\times 1} &= \mathbf{G}^T\tilde{x} = \mathbf{G}^T\mathbf{\Sigma}_{11}^{-1/2}\mathbf{x},\\ \tilde{v}_{q\times 1} &= \mathbf{H}^T\tilde{y} = \mathbf{H}^T\mathbf{\Sigma}_{22}^{-1/2}\mathbf{y},\end{aligned}$$

(similar to pc's) with

$$\text{cov}\left[\begin{pmatrix}\tilde{u}\\ \tilde{v}\end{pmatrix}\right] = \begin{pmatrix}\mathbf{I}_p & \mathbf{R}\\ \mathbf{R}^T & \mathbf{I}_q\end{pmatrix}.$$

Thus, with $\{\mathbf{g}_i\}$ the columns of $\mathbf{G}$ and $\{\mathbf{h}_i\}$ the columns of $\mathbf{H}$, the canonical pairs (maximizing the correlations, and uncorrelated with each other) are, for $i = 1, ..., p$,

$$\left\{\left(\tilde{U}_i, \tilde{V}_i\right)\right\} = \left\{\left(\mathbf{g}_i^T\mathbf{\Sigma}_{11}^{-1/2}\mathbf{x}, \mathbf{h}_i^T\mathbf{\Sigma}_{22}^{-1/2}\mathbf{y}\right)\right\}. \tag{18.3}$$

The i^{th} canonical correlation is

$$\text{corr}\left[\tilde{U}_i, \tilde{V}_i\right] = \rho_i^*,$$

(the singular values of $\tilde{\mathbf{\Sigma}}_{12}$), and

$$\text{var}\left[\tilde{U}_i\right] = \text{var}\left[\tilde{V}_i\right] = 1.$$

- It is often simpler to carry out the discussion in terms of the standardized variables $\tilde{x}$ and $\tilde{y}$, for which the canonical pairs are $\left\{\left(\mathbf{g}_i^T\tilde{x}, \mathbf{h}_i^T\tilde{y}\right)\right\}_{i=1}^p$ and the canonical correlations are the ρ_i^*.

- Note that $\|\mathbf{g}_i\| = \|\mathbf{h}_i\| = 1$, $i = 1, ..., p$. We sometimes work instead with the pairs $\left(\mathbf{a}_i^T\mathbf{x}, \mathbf{b}_i^T\mathbf{y}\right)$ and assume that $\|\mathbf{a}_i\| = \|\mathbf{b}_i\| = 1$, $i = 1, ..., p$. To enforce this, we normalize:

$$U_i = \mathbf{a}_i^T\mathbf{x},\ V_i = \mathbf{b}_i^T\mathbf{y};$$

here (from (18.3))

$$\begin{aligned} \mathbf{a}_i &= \mathbf{\Sigma}_{11}^{-1/2}\mathbf{g}_i \Big/ \left\|\mathbf{\Sigma}_{11}^{-1/2}\mathbf{g}_i\right\|, \\ \mathbf{b}_i &= \mathbf{\Sigma}_{22}^{-1/2}\mathbf{h}_i \Big/ \left\|\mathbf{\Sigma}_{22}^{-1/2}\mathbf{h}_i\right\|. \end{aligned}$$

This does not affect the correlations but

$$\begin{aligned} \text{var}\left[U_i\right] &= 1/\left\|\mathbf{\Sigma}_{11}^{-1/2}\mathbf{g}_i\right\|^2 = 1/\left[\mathbf{G}^T\mathbf{\Sigma}_{11}^{-1}\mathbf{G}\right]_{ii}, \\ \text{var}\left[V_i\right] &= 1/\left\|\mathbf{\Sigma}_{22}^{-1/2}\mathbf{h}_i\right\|^2 = 1/\left[\mathbf{H}^T\mathbf{\Sigma}_{22}^{-1}\mathbf{H}\right]_{ii}. \end{aligned}$$

- Note that

$$\begin{aligned}\sum_{i=1}^{p} \text{var}\left[X_i\right] &= tr\mathbf{\Sigma}_{11} = \sum_{i=1}^{p} \left[\mathbf{G}^T\mathbf{\Sigma}_{11}\mathbf{G}\right]_{ii} \\ &\overset{why?}{\geq} \sum_{i=1}^{p} \text{var}\left[U_i\right], \qquad (18.4)\end{aligned}$$

  (the 'why' is assigned); hence there is no decomposition of the total variance (of $\mathbf{x}$) here.

- There are examples in which the first few canonical variables account for very little variability in each set; e.g.

$$\mathbf{\Sigma} = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 1 & .95 & 0 \\ 0 & .95 & 1 & 0 \\ 0 & 0 & 0 & 100 \end{pmatrix};$$

  Here the highly correlated variables have very little variation: the first canonical pair is (X_2, Y_1) and the second is (X_1, Y_2). Hence we do have equality in (18.4) but var$[U_1] =$ var$[V_1] = 1$ only.

<u>Calculations</u>: We have $p = q = 2$;

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{12} &= \mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2} \\
&= \begin{pmatrix} 1/10 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ .95 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1/10 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 \\ .95 & 0 \end{pmatrix}; \\
\tilde{\mathbf{\Sigma}}_{12}\tilde{\mathbf{\Sigma}}_{21} &= \begin{pmatrix} 0 & 0 \\ 0 & (.95)^2 \end{pmatrix}; \\
\rho_1^* &= .95,\ \rho_2^* = 0; \\
\mathbf{g}_1 &= \binom{0}{1},\ \mathbf{g}_2 = \binom{1}{0}; \\
\tilde{\mathbf{\Sigma}}_{21}\tilde{\mathbf{\Sigma}}_{12} &= \begin{pmatrix} (.95)^2 & 0 \\ 0 & 0 \end{pmatrix}; \\
\mathbf{h}_1 &= \binom{1}{0},\ \mathbf{h}_2 = \binom{0}{1}; \\
\left(\tilde{U}_1, \tilde{V}_1\right) &= \left(\tilde{X}_2, \tilde{Y}_1\right) \propto (X_2, Y_1) = (U_1, V_1); \\
\left(\tilde{U}_2, \tilde{V}_2\right) &= \left(\tilde{X}_1, \tilde{Y}_2\right) \propto (X_1, Y_2) = (U_2, V_2).
\end{aligned}$$

19. Canonical correlations II

- **Sample canonical correlations**: The sample canonical variables are obtained by applying the population method to $\mathbf{S}$ (one can use $\mathbf{R}$, but then all interpretations are in terms of linear combinations of <u>standardized</u> variables).

- To compute: obtain $\mathbf{S}_{11}^{-1/2}$, $\mathbf{S}_{22}^{-1/2}$ and then $\tilde{S}_{12} = \mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}$. Use

  $$\mathtt{s} = \mathtt{svd}(\tilde{S}_{12},\ \ \mathtt{nu} = \mathtt{p},\ \ \mathtt{nv} = \mathtt{q})$$

  to get

  $$\mathbf{G} = \mathtt{s\$u}, \mathbf{H} = \mathtt{s\$v} \text{ and } \left(\hat{\rho}_1^*, \cdots, \hat{\rho}_p^*\right) = \mathtt{s\$d}.$$

  Then the normalized rows of $\tilde{A} = \mathbf{G}^T\mathbf{S}_{11}^{-1/2}$ and $\tilde{B} = \mathbf{H}^T\mathbf{S}_{22}^{-1/2}$ are the $\left(\mathbf{a}_i^T, \mathbf{b}_i^T\right)$, and $(U_i, V_i) = \left(\mathbf{a}_i^T\mathbf{x}, \mathbf{b}_i^T\mathbf{y}\right)$ are the canonical pairs.

- Example 10.1 – see web site.

```
> round(cbind(A, B),2)
# coefficients of U and V
      [,1]  [,2]  [,3]  [,4]
[1,] -0.95 -0.31 -0.59 -0.80
[2,] -0.54  0.84 -0.77  0.63

> round(1/diag(Atilde%*%t(Atilde)),2) # var of U's
[1] 1.23 0.64
> round(1/diag(Btilde%*%t(Btilde)),2) # var of V's
[1] 1.19 0.80

> # Canonical correlations:
> round(diag(R),2)
[1] 0.74 0.03
```

Interpretation:

$$\begin{aligned}
(U_1, V_1) &= (-.95X_1 - .31X_2, -.59Y_1 - .80Y_2),\\
(U_2, V_2) &= (-.54X_1 + .84X_2, -.77Y_1 + .63Y_2);\\
|\text{corr}\,[U_1, V_1]| &= .74, \quad |\text{corr}\,[U_2, V_2]| = .03.
\end{aligned}$$

The variances are

$$\begin{aligned} \text{var}[U_1] &= 1.23, \text{var}[V_1] = 1.19; \\ \text{var}[U_2] &= .64,\ \text{var}[V_2] = .80; \end{aligned}$$

with

$$\text{var}[U_1] + \text{var}[U_2] = 1.87 < 2 = \text{var}[X_1] + \text{var}[X_2].$$

- Relationship to other correlations.
  - When $p = q = 1$ all 'linear combinations' are merely multiples, and the correlation is unaffected – there is only one canonical pair and its correlation is (in absolute value) the ordinary correlation between X and Y.
  - When $p = 1$ the first canonical correlation is
    $$\rho_1^* = \max_{\mathbf{b}} \text{corr}\left[X, \mathbf{b}^T\mathbf{y}\right],$$
    the multiple correlation coefficient (Lecture 13) between X and $\mathbf{y}$.

- If $\mathbf{\Sigma}_{12} = \mathbf{0}$ then all $\rho_i^* = 0$. We can test this hypothesis; if rejected we might test the hypothesis that only the first k are significantly non-zero. Large sample likelihood ratio tests with Bartlett corrections:

  - $H : \mathbf{\Sigma}_{12} = \mathbf{0}$. Reject for large values of

    $$\begin{aligned} -2 \log \Lambda &\overset{how?}{=} n \log \left(\frac{|\mathbf{S}_{11}|\,|\mathbf{S}_{22}|}{|\mathbf{S}|} \right) \\ &\overset{assigned}{=} -n \log \prod_{i=1}^{p} \left(1 - \hat{\rho}_i^{*2} \right) \xrightarrow{L} \chi^2_{pq}. \end{aligned} \tag{19.1}$$

    The Bartlett correction replaces n by $n - 1 - \frac{p+q+1}{2}$.

  - $H : \rho_1^* \geq \cdots \geq \rho_k^* > 0 = \rho_{k+1}^* = \cdots = \rho_p^*$:

    $$-\left(n - 1 - \frac{p+q+1}{2} \right) \log \prod_{i=k+1}^{p} \left(1 - \hat{\rho}_i^{*2} \right) \xrightarrow{L} \chi^2_{(p-k)(q-k)}.$$

- Example 10.5, 10.8. Two sets of variables – ‘job characteristics’ and ‘measures of job satisfaction’ ($p = 5$, $q = 7$):

$$\begin{aligned}\mathbf{x} &= \begin{pmatrix}\text{feedback}\\ \text{task significance}\\ \text{task variety}\\ \text{task identity}\\ \text{autonomy}\end{pmatrix},\\ \mathbf{y} &= \begin{pmatrix}\text{supervisor satisf'n,}\\ \text{career-future satisf'n}\\ \text{financial satisf'n}\\ \text{workload satisf'n}\\ \text{company identification}\\ \text{kind-of-work satisf'n}\\ \text{general satisf'n}\end{pmatrix}.\end{aligned}$$

See code on web site – the coefficients of the first two canonical pairs are

$$\left|\begin{array}{lcc} & \hat{\mathbf{a}}_1 & \hat{\mathbf{a}}_2 \\ X_1 & 0.62 & 0.25 \\ X_2 & 0.29 & -0.48 \\ X_3 & 0.25 & -0.61 \\ X_4 & -0.03 & 0.25 \\ X_5 & 0.68 & 0.52 \end{array}\right|, \left|\begin{array}{lcc} & \hat{\mathbf{b}}_1 & \hat{\mathbf{b}}_2 \\ Y_1 & 0.55 & -0.07 \\ Y_2 & 0.27 & 0.37 \\ Y_3 & -0.05 & -0.08 \\ Y_4 & 0.03 & 0.79 \\ Y_5 & 0.38 & -0.09 \\ Y_6 & 0.67 & -0.47 \\ Y_7 & -0.14 & -0.03 \end{array}\right|.$$

So U_1 is a ‘feedback’ and ‘autonomy’ variable (X_1 and X_5; ‘task’ variables don’t seem to be so important) while V_1 stresses satisfaction through ‘supervisor’ and ‘kind of work’ (Y_1 and Y_6).

- To augment this interpretation one can look at the correlations between the canonical variables and the original ones. If $\tilde{A}$ and $\tilde{B}$ have the coefficients $\tilde{a}_i^T$ and $\tilde{b}_i^T$ of $\tilde{U}_i$ and $\tilde{V}_i$ as their rows (these aren’t normalized, but that doesn’t matter), then the estimates are

$$\begin{aligned}
\text{corr}\left[\hat{U}_i, X_j\right] &= \text{corr}\left[\tilde{U}_i, X_j\right] \\
&= \text{corr}\left[\tilde{a}_i^T \mathbf{x}, \mathbf{e}_j^T \mathbf{x}\right] = \frac{\tilde{a}_i^T \mathbf{S}_{11} \mathbf{e}_j}{\sqrt{\mathbf{e}_j^T \mathbf{S}_{11} \mathbf{e}_j}} \\
&= \left[\left(\tilde{A}^T \mathbf{S}_{11}\right)\left(diag\left(\mathbf{S}_{11}\right)\right)^{-1/2}\right]_{ij};
\end{aligned}$$

similarly

$$\begin{aligned}
\text{corr}\left[\hat{U}_i, Y_j\right] &= \left[\left(\tilde{A}^T \mathbf{S}_{12}\right)\left(diag\left(\mathbf{S}_{22}\right)\right)^{-1/2}\right]_{ij}, \\
\text{corr}\left[\hat{V}_i, X_j\right] &= \left[\left(\tilde{B}^T \mathbf{S}_{21}\right)\left(diag\left(\mathbf{S}_{11}\right)\right)^{-1/2}\right]_{ij}, \\
\text{corr}\left[\hat{V}_i, Y_j\right] &= \left[\left(\tilde{B}^T \mathbf{S}_{22}\right)\left(diag\left(\mathbf{S}_{22}\right)\right)^{-1/2}\right]_{ij}.
\end{aligned}$$

For the first canonical pair this gives the correlations

$$\left|\begin{array}{lcc}
 & \left(U_1, X_j\right) & \left(V_1, X_j\right) \\
X_1 & 0.83 & 0.46 \\
X_2 & 0.73 & 0.42 \\
X_3 & 0.75 & 0.61 \\
X_4 & 0.62 & 0.34 \\
X_5 & 0.86 & 0.48
\end{array}\right|, \left|\begin{array}{lcc}
 & \left(U_1, Y_j\right) & \left(V_1, Y_j\right) \\
Y_1 & 0.42 & 0.76 \\
Y_2 & 0.36 & 0.64 \\
Y_3 & 0.21 & 0.39 \\
Y_4 & 0.21 & 0.38 \\
Y_5 & 0.36 & 0.65 \\
Y_6 & 0.45 & 0.80 \\
Y_7 & 0.28 & 0.50
\end{array}\right|.$$

Here we see that U_1 has about the same correlation with all of the X_i (although still greatest with X_1 and X_5), perhaps contradicting the interpretation above based just on the coefficients of U_1. But the interpretation of V_1 as reflecting satisfaction through Y_1 and Y_6 continues to hold.

- Test $H_k : \rho_1^* \geq \cdots \geq \rho_k^* > 0 = \rho_{k+1}^* = \cdots = \rho_p^*$ – that only k of the canonical correlations are significant – for $k = 0$, then $k = 1$, etc. The p-values of each are

$$p_k = P\left(\begin{array}{c} \chi^2_{(p-k)(q-k)} > \\ -\left(n-1-\frac{p+q+1}{2}\right)\log\prod_{i=k+1}^{p}\left(1-\hat{\rho}_i^{*2}\right)\end{array}\right).$$

k	χ^2	df	p_k
0	346.68	35	0.00
1	62.38	24	0.00
2	17.72	15	0.28
3	6.61	8	0.58
4	2.55	3	0.47

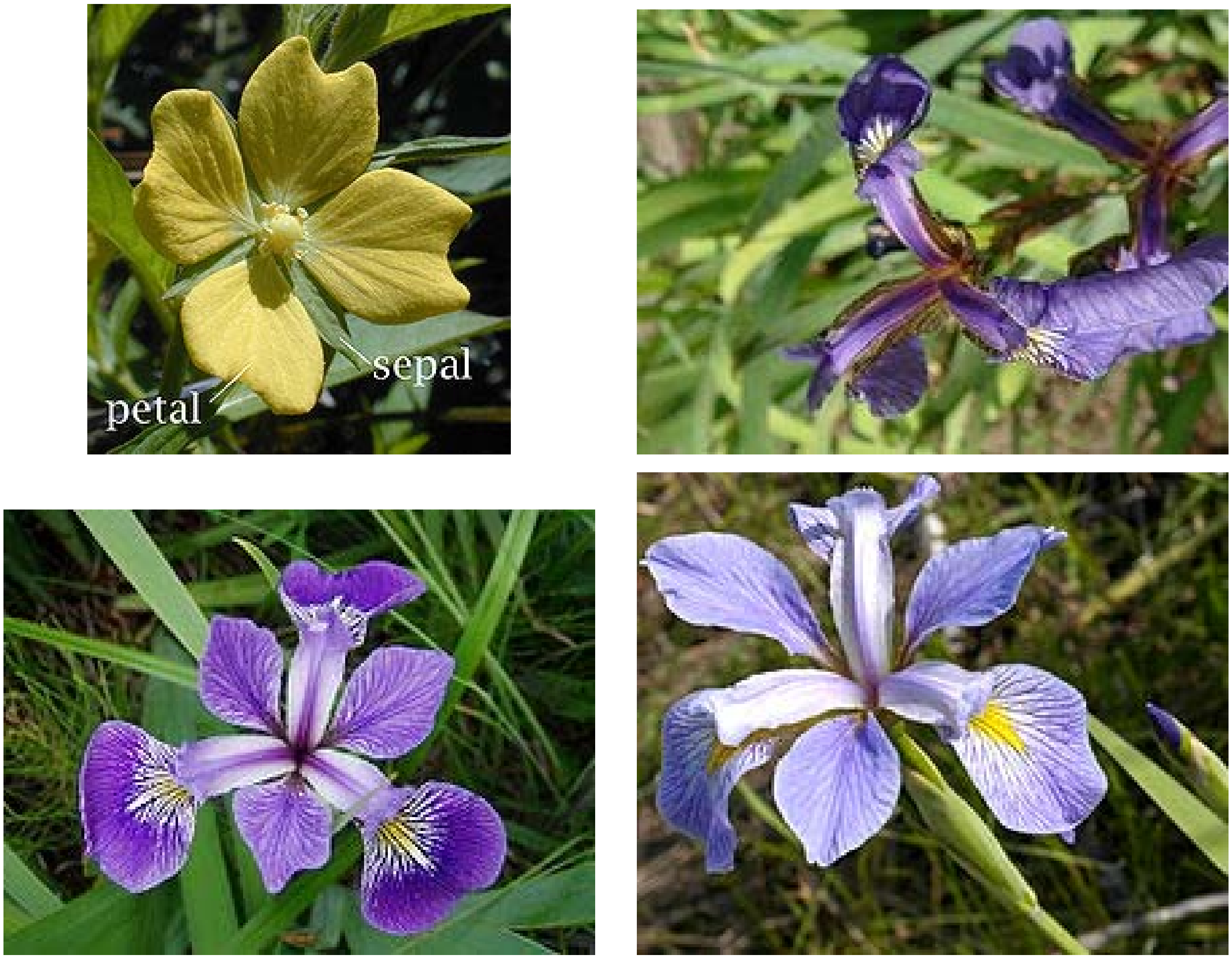
Interpretation?

Part IV

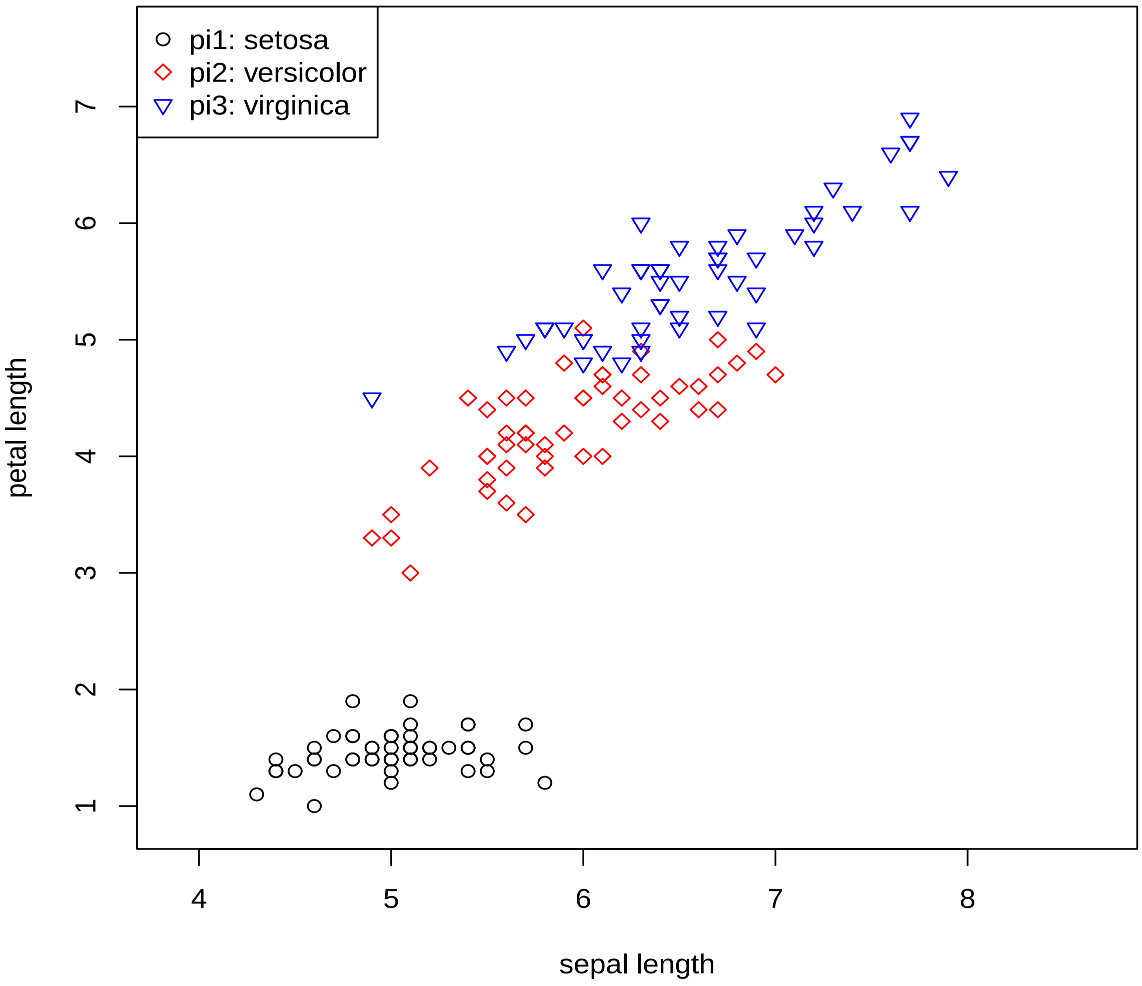
CLASSIFICATION AND GROUPING

20. Discrimination and classification: strategies

- Example – Observations made on four attributes (sepal length/width, petal length/width) of each of three types of irises. Top right: Setosa, Bottom left: Versicolor, Bottom right: Virginica.



- Two of these attributes:



- Classification problem: For a class variable $Y \in \{1, 2, ..., J\}$ and data $\mathbf{x} \in \mathbb{R}^p$ develop a function $\delta(\mathbf{x}) = y$; classify $\mathbf{x}$ as having arisen from population π_j if $y = j$.

- Distributions of interest:

  – conditional (posterior) class distribution $p(y|\mathbf{x})$,

  – marginal (prior) class distribution $p(y)$,

  – conditional feature density $f(\mathbf{x}|y)$,

  – marginal density $f(\mathbf{x}) = \sum_y f(\mathbf{x}|y)\,p(y)$.

- We incur a loss if $Y = y$ but we incorrectly classify as $j \neq y$. In general, define $L(j, y)$ to be the loss when $Y = y$ but $\delta(\mathbf{x}) = j$. Often ("equal losses")

$$L(j, y) = I(y \neq j).$$

The 'risk' associated with the rule δ is the expected loss:

$$\begin{aligned} R_\delta &= E\left[L\left(\delta\left(\mathbf{x}\right), Y\right)\right] \\ &= E\left[E\left[L\left(\delta\left(\mathbf{x}\right), Y\right)|\mathbf{x}\right]\right], \end{aligned}$$

where now the outer expectation if over the distribution of $\mathbf{x}$, the inner – $h\left(\mathbf{x}\right) = E\left[L\left(\delta\left(\mathbf{x}\right), Y\right)|\mathbf{x}\right]$ – is over the conditional distribution of Y, given $\mathbf{x}$. This is the Double Expectation Theorem; also a version of 'Bayes' rule'.

- Clearly, the risk is minimized if $h\left(\mathbf{x}\right)$ is minimized for each $\mathbf{x}$. For equal losses,

$$\begin{aligned} h\left(\mathbf{x}\right) &= P\left(Y \neq \delta\left(\mathbf{x}\right)|\mathbf{x}\right) = 1 - P\left(Y = \delta\left(\mathbf{x}\right)|\mathbf{x}\right) \\ &= 1 - p\left(y|\mathbf{x}\right)\big|_{y=\delta(\mathbf{x})} \end{aligned}$$

and so we choose

$$\delta\left(\mathbf{x}\right) = \arg\max_y p\left(y|\mathbf{x}\right).$$

Thus in this case Bayes' rule classifies $\mathbf{x}$ in the class with the largest posterior probability. Since

$$p\left(y|\mathbf{x}\right) = \frac{f\left(\mathbf{x}|y\right)p\left(y\right)}{f\left(\mathbf{x}\right)},$$

in the case of 'equal priors' – $p(y) = 1/J$ for each y – this reduces to the rule

$$\delta(\mathbf{x}) = \arg\max_y f(\mathbf{x}|y).$$

- We can write the risk as

$$\begin{aligned} R_\delta &= E[L(\delta(\mathbf{x}), Y)] = E[E[L(\delta(\mathbf{x}), Y)|Y]] \\ &= \sum_y p(y) \sum_j L(j, y) P(\delta(\mathbf{x}) = j|Y = y); \end{aligned}$$

  the function $P(\delta(\mathbf{x}) = j|Y = y)$ is the 'misclassification probability' (when $y \neq j$).

- There are several strategies for constructing classification rules.
  Strategy 1: Start with parametric models for $f(\mathbf{x}|y)$.
  For instance if this is the $N(\mu_y, \mathbf{\Sigma})$ density, then

$$\begin{aligned} p(y|\mathbf{x}) &\propto \exp\left\{\log p(y) - \frac{1}{2}(\mathbf{x} - \mu_y)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu_y)\right\} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (20.1) \\ &\propto \exp\left\{\begin{array}{c} \log p(y) - \frac{1}{2}\mu_y^T \mathbf{\Sigma}^{-1}\mu_y \\ +\mu_y^T \mathbf{\Sigma}^{-1}\mathbf{x} \end{array}\right\} \qquad (20.2) \\ &= \exp\left\{\alpha_y + \beta_y^T \mathbf{x}\right\}, \text{ say.} \end{aligned}$$

Thus

$$p\left(y|\mathbf{x}\right)=\frac{\exp\left\{\alpha_y+\beta_y^T\mathbf{x}\right\}}{\sum_j\exp\left\{\alpha_j+\beta_j^T\mathbf{x}\right\}}.$$

(With equal priors we can ignore the $p\left(y\right)$.) Bayes' rule is to classify $\mathbf{x}$ into the class for which the numerator is a maximum. For instance if there are only two classes we classify as '1' rather than '2' if

$$\begin{aligned} 0 &< \left(\beta_1-\beta_2\right)^T\mathbf{x}+\left(\alpha_1-\alpha_2\right) \\ &= \left(\mu_1-\mu_2\right)^T\mathbf{\Sigma}^{-1}\mathbf{x}+\log\frac{p(1)}{p(2)} \\ &\quad -\frac{1}{2}\left(\mu_1-\mu_2\right)^T\mathbf{\Sigma}^{-1}\left(\mu_1+\mu_2\right)^T \\ &= \left(\mu_1-\mu_2\right)^T\mathbf{\Sigma}^{-1}\left[\mathbf{x}-\frac{1}{2}\left(\mu_1+\mu_2\right)\right] \\ &\quad +\log\frac{p(1)}{p(2)}. \end{aligned}$$

- Geometrically, we see which side of a hyperplane $\mathbf{x}$ lies on, and classify accordingly – 'linear discriminant analysis'. Sometimes this entire quantity is divided by the norm of (the estimate of) $\mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$, so that the vector of coefficients of $\mathbf{x}$ has a norm of 1; this ('scaling' – p. 589 in text) can make the output easier to interpret.

  - When the covariances are not equal, (20.2) must contain $-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_y^{-1}\mathbf{x}$, leading to

    $$p(y|\mathbf{x}) = \frac{\exp\left\{\alpha_y + \beta_y^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_y^{-1}\mathbf{x}\right\}}{\sum_j \exp\left\{\alpha_j + \beta_j^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_y^{-1}\mathbf{x}\right\}},$$

    with $\mathbf{\Sigma}$ replaced by $\mathbf{\Sigma}_y$ in the definitions of α_y and β_y – 'quadratic discriminant analysis':

    $$\delta(\mathbf{x}) = \arg\max_y \left[\alpha_y + \beta_y^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_y^{-1}\mathbf{x}\right].$$

- These formulae obscure the simplicity of the procedures. Define 'Mahalanobis distances'

  $$\|\mathbf{x}-\mu\|_{\mathbf{\Sigma}}=\sqrt{(\mathbf{x}-\mu)^T\,\mathbf{\Sigma}^{-1}\,(\mathbf{x}-\mu)}.$$

  Then (from (20.2)), with priors $\left\{p_j\right\}_{j=1}^J$ the rule is to classify into $N\left(\mu_{j^*},\mathbf{\Sigma}\right)$, if

  $$j^*=\arg\min_j\left\{\left\|\mathbf{x}-\mu_j\right\|_{\mathbf{\Sigma}}^2-2\log p_j\right\};$$

  the quadratic rule is to classify into $N\left(\mu_{j^*},\mathbf{\Sigma}_{j^*}\right)$, if

  $$j^*=\arg\min_j\left\{\left\|\mathbf{x}-\mu_j\right\|_{\mathbf{\Sigma}_j}^2-2\log p_j\right\}.$$

  With equal priors these are

  $$j^*=\arg\min_j\left\|\mathbf{x}-\mu_j\right\|_{\mathbf{\Sigma}}^2$$

  and

  $$j^*=\arg\min_j\left\|\mathbf{x}-\mu_j\right\|_{\mathbf{\Sigma}_j}^2$$

  respectively.

- To apply these methods one gathers a 'training sample', for which the correct classes are known; this results in samples from which the μ_y can be estimated (by the sample averages) and Σ or Σ_y can be estimated by the pooled or individual sample covariances. The prior $p\left(y\right)$ is generally estimated by the proportion of the members of the training sample belonging to class y. For instance with equal priors and equal covariances, we classify $\mathbf{x}$ into Population 1 if

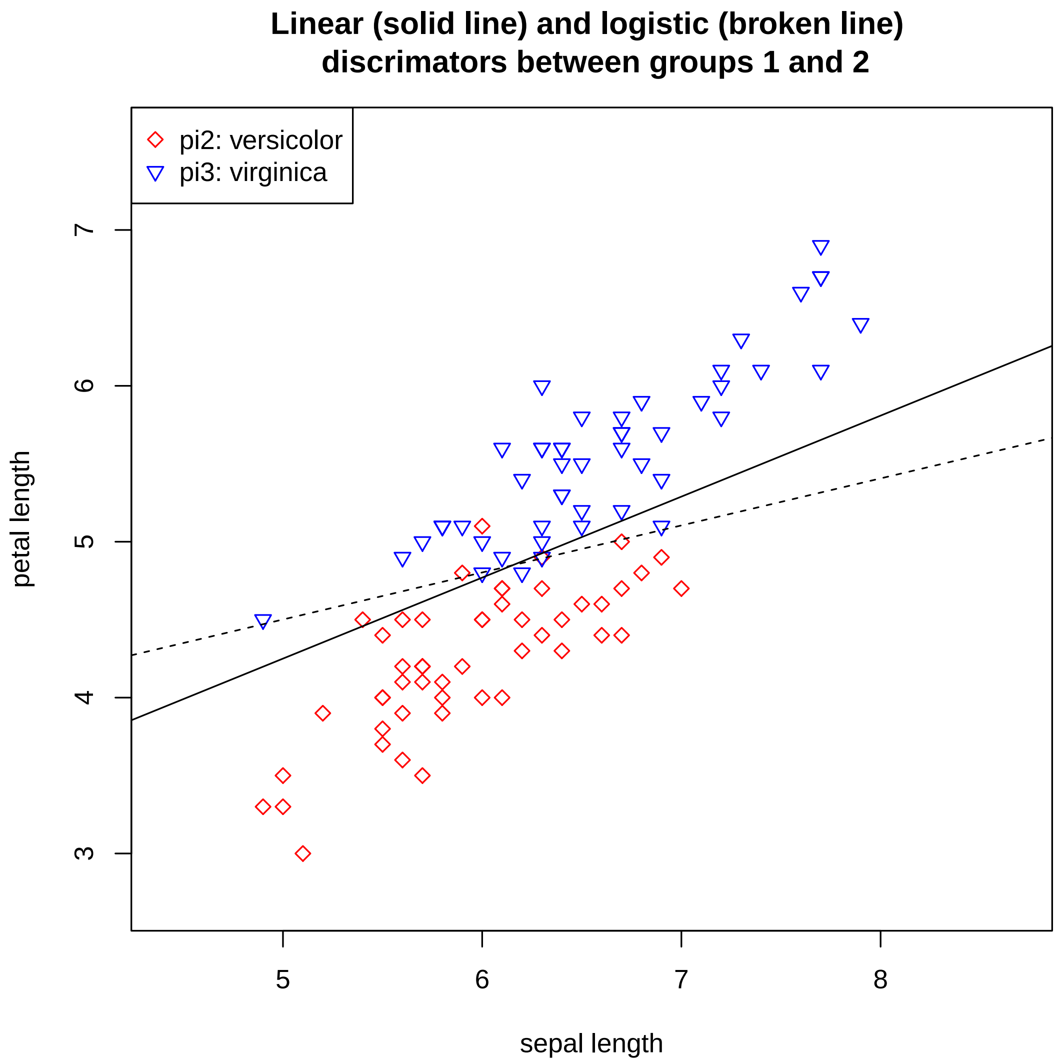
$$\left\|\mathbf{x}-\bar{\mathbf{x}}_1\right\|^2_{\mathbf{S}_{pooled}} \leq \left\|\mathbf{x}-\bar{\mathbf{x}}_2\right\|^2_{\mathbf{S}_{pooled}};$$

this becomes the linear discrimination rule

$$0 < \left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)^T \mathbf{S}^{-1}_{pooled}\left[\mathbf{x}-\frac{1}{2}\left(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2\right)\right]. \tag{20.3}$$

  - The R functions `lda` and `qda`, in the MASS library, carry out linear and quadratic discrimination. The output is called by `predict()` to then classify new observations. By default the prior probabilities are estimated by the sample proportions.

– For the Iris data, *after dropping group 3*, the lda results in the following plot (solid line; code on web site).



- Strategy 2: Consider models for $p\left(y|\mathbf{x}\right)$ directly – logistic regression. First suppose there are only 2 classes, so that Y is Bernoulli, with
$$P\left(Y=1|\mathbf{x}\right)=p\left(1|\mathbf{x}\right)=1-p\left(0|\mathbf{x}\right).$$
Model the 'logits' linearly:
$$\log\frac{p\left(1|\mathbf{x}\right)}{1-p\left(1|\mathbf{x}\right)}=\alpha+\beta^T\mathbf{x}.$$
Then
$$p\left(1|\mathbf{x}\right)=\frac{e^{\alpha+\beta^T\mathbf{x}}}{1+e^{\alpha+\beta^T\mathbf{x}}}$$
and, with $y_i\in\{0,1\}=I\left(\text{obs'n i}\in\text{class 1}\right)$, the likelihood function is
$$\begin{aligned}L &= \prod_{i=1}^{n}p\left(y_i|\mathbf{x}_i\right)\\ &= \prod_{i=1}^{n}\left(p\left(1|\mathbf{x}_i\right)\right)^{y_i}\left(1-p\left(1|\mathbf{x}_i\right)\right)^{1-y_i}\\ &= \prod_{i=1}^{n}\left(\frac{e^{\alpha+\beta^T\mathbf{x}_i}}{1+e^{\alpha+\beta^T\mathbf{x}_i}}\right)^{y_i}\left(\frac{1}{1+e^{\alpha+\beta^T\mathbf{x}_i}}\right)^{1-y_i}\\ &= \prod_{i=1}^{n}\frac{e^{y_i\left(\alpha+\beta^T\mathbf{x}_i\right)}}{\left(1+e^{\alpha+\beta^T\mathbf{x}_i}\right)}.\end{aligned}$$

- Given a two-group (labelled 1 and 0) training sample $\{y_i, \mathbf{x}_i\}$ the mles $\hat{\alpha}, \hat{\beta}$ are computed. When future observations with features $\mathbf{x}$ are made we estimate
$$\hat{p}\left(1|\mathbf{x}\right) = \frac{e^{\hat{\alpha}+\hat{\beta}^T\mathbf{x}}}{1+e^{\hat{\alpha}+\hat{\beta}^T\mathbf{x}}}$$
and classify the observation into Class 1 if $\hat{p}\left(1|\mathbf{x}\right) > \hat{p}\left(0|\mathbf{x}\right)$, i.e. if $\hat{\alpha} + \hat{\beta}^T\mathbf{x} > 0$, and into Class 0 otherwise.

  - The R function `fit = glm(class ~X, family = binomial())`, with 'class' the $0-1$ vector of group memberships and $\mathbf{X}$ the training sample does the fitting for two classes.

  - Then `predict(fit, as.data.frame(X))` returns the values of $\hat{\alpha} + \hat{\beta}^T\mathbf{x}$.

  - Example on web site; broken line in the last plot.

– If there are J populations and data $\left\{\mathbf{x}_{jk}\right\}_{k=1}^{n_j}$ arising from population $j \in \{1, ..., J\}$, then the likelihood is

$$L = \prod_{j=1}^{J} \prod_{k=1}^{n_j} p\left(j|\mathbf{x}_{jk}\right),$$

with

$$\begin{aligned} p\left(j|\mathbf{x}\right) &= \frac{e^{\alpha_j + \beta_j^T \mathbf{x}}}{1 + \sum_{j=1}^{J-1} e^{\alpha_j + \beta_j^T \mathbf{x}}}, \ j = 1, ..., J-1, \\ p\left(J|\mathbf{x}\right) &= \frac{1}{1 + \sum_{j=1}^{J-1} e^{\alpha_j + \beta_j^T \mathbf{x}}} \\ &= 1 - p_1 - \cdots - p_{J-1}. \end{aligned}$$

Classify an observation with features $\mathbf{x}$ into class

$$j^* = \text{arg max}\, \hat{p}\left(j|\mathbf{x}\right).$$

Evidently there is no inbuilt R function for multivariate logistic regression; I have written one and put it on the web site.

21. Discrimination and classification: assessment

- Suppose there are only two populations, with priors $p(1)$ and $p(2) = 1 - p(1)$, and densities $f(\mathbf{x}|j)$, $j = 1, 2$. For a classifier $\delta(\mathbf{x})$ define regions

  $$R_j = \{\mathbf{x}|\delta(\mathbf{x}) = j\}.$$

  (Then $R_2 = \Omega \backslash R_1$.) The conditional probabilities of misclassification $P(j|y) = P(\delta(\mathbf{x}) = j|y)$ are

  $$\begin{aligned} P(1|2) &= \int_{R_1} f(\mathbf{x}|2)\, d\mathbf{x} = 1 - P(2|2), \\ P(2|1) &= \int_{R_2} f(\mathbf{x}|1)\, d\mathbf{x} = 1 - P(1|1). \end{aligned}$$

  Then the overall probabilities of correctly or incorrectly classifying are as follows.

	Classify as:	
when:	Pop'n 1	Pop'n 2
True pop'n is 1	$p(1)P(1\|1)$	$p(1)P(2\|1)$
True pop'n is 2	$p(2)P(1\|2)$	$p(2)P(2\|2)$

- Suppose that there are costs $c\,(1|2)$ and $c(2|1)$ associated with misclassifying an object. The 'expected cost of misclassification' is then

  $$ECM = c(2|1)p(1)P\,(2|1) + c\,(1|2)\,p(2)P\,(1|2)\,.$$

  A reasonable strategy is to choose $\delta\,(\mathbf{x})$ so as to minimize the ECM. This results (assigned; proof is very much like that of the Neyman-Pearson Lemma) in

  $$R_1 = \left\{ \frac{f\,(\mathbf{x}|1)}{f\,(\mathbf{x}|2)} \geq \frac{c\,(1|2)\,p(2)}{c(2|1)p(1)} \right\} = R_2^c. \tag{21.1}$$

  - With equal costs this is

    $$\delta\,(\mathbf{x}) = \text{arg max}\, f\,(\mathbf{x}|y)\,p(y) = \text{arg max}\, p(y|\mathbf{x}),$$

    i.e. is Bayes' rule.

  - The 'total probability of misclassification' is

    $$TPM = p(1)P\,(2|1) + p(2)P\,(1|2)\,,$$

    i.e. is ECM with equal (to 1) costs; thus Bayes' rule also minimizes TPM.

- For J populations this extends to

  $$\delta(\mathbf{x}) = \arg\min_k \sum_{\substack{j=1 \\ j\neq k}}^{J} f(\mathbf{x}|j)\, p(j) c(k|j)$$

  as the minimizer of ECM. (The sum is the integrand of the EC of misclassifying as 'k' – how?) With equal costs this is

  $$\delta(\mathbf{x}) = \arg\max_k f(\mathbf{x}|k)\, p(k) = \arg\max p(y|\mathbf{x});$$

  again this is Bayes' rule.

- Optimum error rate (OER) = minimum possible TPM. Example – Discriminate between two normal populations; equal covariances, means μ_1 and μ_2. Assume equal costs and priors, so that

  $$R_1 = \left\{ (\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1} \left[\mathbf{x} - \frac{1}{2}(\mu_1 + \mu_2) \right] \geq 0 \right\}$$

  and

  $$OER = \frac{1}{2} P(\mathbf{x} \in R_1|2) + \frac{1}{2} P(\mathbf{x} \in R_2|1).$$

Put $Y = (\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}\mathbf{x}$. This is normal, variance $(\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1} (\mu_1 - \mu_2) \overset{def}{=} \Delta^2$ and mean

$$\begin{aligned} \nu_1 &= E[Y|1] = (\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}\mu_1, \\ \nu_2 &= E[Y|2] = (\mu_1 - \mu_2)^T \boldsymbol{\Sigma}^{-1}\mu_2, \end{aligned}$$

with $\nu_1 - \nu_2 = \Delta^2$. We have

$$\begin{aligned} P(\mathbf{x} \in R_1|2) &= P\left(Y \geq \frac{\nu_1 + \nu_2}{2}|2\right) \\ &= P\left(\frac{Y - \nu_2}{\Delta} \geq \frac{\Delta}{2}\right) \\ &= \Phi\left(-\frac{\Delta}{2}\right). \end{aligned}$$

Similarly $P(\mathbf{x} \in R_2|1) = \Phi(-\Delta/2)$ and so

$$OER = \Phi\left(-\frac{\Delta}{2}\right) = \Phi\left(-\frac{1}{2}\|\mu_1 - \mu_2\|_{\boldsymbol{\Sigma}}\right).$$

For small Δ this is almost .5.

- One way to estimate TPM is by the ‘actual error rate’

$$AER = p(1)\int_{\hat{R}_2} f(\mathbf{x}|1)\, d\mathbf{x} + p(2)\int_{\hat{R}_1} f(\mathbf{x}|2)\, d\mathbf{x},$$

where $\hat{R}_1$ and $\hat{R}_2$ are the regions computed from the training sample. Requires the densities to be known.

- Another is to apply the classifier to the training sample and compute the 'apparent error rate' APER – the fraction of items misclassified. This is based on the 'confusion matrix' with elements $m_{ij} = \#$ of items from pop'n i classified into pop'n j. Then
$$APER = \frac{\sum_{i \neq j} m_{ij}}{n}.$$
Example in code on web site. This method is clearly biased (why?).

- **Lachenbruch's holdout method**. Omit one observation from the training sample; develop the classifier from the the remainder of the sample; classify the observation which was held out. Repeat n times; estimate the error rate from the resulting confusion matrix. Examples on web site.

```
Linear discrimination: APER = 0.02
Confusion matrix is
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   48    2
[3,]    0    1   49

Quadratic discrimination: APER = 0.02
Confusion matrix is
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   48    2
[3,]    0    1   49

Logistic discrimination: APER = 0.013
Confusion matrix is
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   49    1
[3,]    0    1   49
```

```
Linear discrimination: estimated rate = 0.02
Confusion matrix is
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   48    2
[3,]    0    1   49

Quadratic discrimination: estimated rate = 0.027
Confusion matrix is
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   47    3
[3,]    0    1   49

Logistic discrimination: estimated rate = 0.02
Confusion matrix is
     [,1] [,2] [,3]
[1,]   50    0    0
[2,]    0   48    2
[3,]    0    1   49
```

22. Discrimination and classification: Fisher's methods

- Example (from STAT 512): Suppose we are given lengths and widths of n prehistoric skulls, of type A or B (the "training sample"). We know that n_1 of these, say $\mathbf{x}_1, ...\mathbf{x}_{n_1}$, are of type A, and $n_2 = n - n_1$, say $\mathbf{y}_1, ...\mathbf{y}_{n_2}$, are of type B. Now we find a new skull, with length and width the components of $\mathbf{z}$. We are to classify it as A or B. (Others applications: rock samples in geology, risk data in an actuarial analysis, etc.).

- Fisher's approach to problems of this type was to reduce them to simpler, univariate problems.

- Who was Fisher?

Sir Ronald Aylmer Fisher FRS (17 February 1890 – 29 July 1962) was an English statistician, evolutionary biologist, eugenicist and geneticist. Known as the “father of modern statistics”. Anders Hald called him “a genius who almost single-handedly created the foundations for modern statistical science” while Richard Dawkins named him “the greatest biologist since Darwin”.

- Define $u_i = \boldsymbol{\alpha}^T \mathbf{x}_i$, $v_i = \boldsymbol{\alpha}^T \mathbf{y}_i$ for some vector $\boldsymbol{\alpha}$. Put $w = \boldsymbol{\alpha}^T \mathbf{z}$ and classify new skull as A if $|w - \bar{u}| < |w - \bar{v}|$.

- Choose $\boldsymbol{\alpha}$ for "maximal separation": $|\bar{u} - \bar{v}|$ should be large relative to the underlying variation. Put

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1} \sum (u_i - \bar{u})^2 = \frac{1}{n_1 - 1} \sum \left(\boldsymbol{\alpha}^T (\mathbf{x}_i - \bar{\mathbf{x}}) \right)^2 \\ &= \frac{1}{n_1 - 1} \boldsymbol{\alpha}^T \sum (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{S}_1 \boldsymbol{\alpha} \end{aligned}$$

and similarly define s_2^2 as the variation in the other sample. Choose $\boldsymbol{\alpha}$ to maximize

$$\begin{aligned} & \frac{(\bar{u} - \bar{v})^2}{\left[(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 \right] / (n - 2)} \\ = \ & \frac{\boldsymbol{\alpha}^T (\bar{\mathbf{x}} - \bar{\mathbf{y}}) (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{S} \boldsymbol{\alpha}}, \qquad (22.1) \end{aligned}$$

where

$$\mathbf{S} = \frac{(n_1 - 1) \mathbf{S}_1 + (n_2 - 1) \mathbf{S}_2}{n - 2} = \mathbf{S}_{pooled}.$$

- Note that if $\mathbf{x} \sim N_p(\mu_1, \mathbf{\Sigma})$ and $\mathbf{y} \sim N_p(\mu_2, \mathbf{\Sigma})$ then

  $$\bar{u} - \bar{v} \sim N\left(\boldsymbol{\alpha}^T(\mu_1 - \mu_2), \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\alpha}^T\mathbf{\Sigma}\boldsymbol{\alpha}\right),$$

  and $E\left[\mathbf{S}_{pooled}\right] = \mathbf{\Sigma}$. But the method does not depend on any of this.

- Put $\boldsymbol{\beta} = \mathbf{S}^{1/2}\boldsymbol{\alpha}$, $\boldsymbol{\alpha} = \mathbf{S}^{-1/2}\boldsymbol{\beta}$ so (22.1) is

  $$\frac{\boldsymbol{\beta}^T\mathbf{S}^{-1/2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T\mathbf{S}^{-1/2}\boldsymbol{\beta}}{\boldsymbol{\beta}^T\boldsymbol{\beta}},$$

  which is a maximum if $\boldsymbol{\beta}$ is the eigenvector corresponding to

  $$ch_{\max}\mathbf{S}^{-1/2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T\mathbf{S}^{-1/2} = ch_{\max}\mathbf{b}\mathbf{b}^T,$$

  where $\mathbf{b} = \mathbf{S}^{-1/2}(\bar{\mathbf{x}} - \bar{\mathbf{y}})$. Note $\mathbf{b}\mathbf{b}^T$ has rank 1, hence has 1 non-zero eigenvalue. This is necessarily the eigenvalue of $\mathbf{b}^T\mathbf{b}$, i.e. is

  $$\lambda = \mathbf{b}^T\mathbf{b} = (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T\mathbf{S}^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) \overset{def}{=} D^2.$$

  Now solve $\mathbf{b}\mathbf{b}^T\boldsymbol{\beta} = \lambda\boldsymbol{\beta}$, i.e.

  $$\mathbf{b}\mathbf{b}^T\boldsymbol{\beta} = \boldsymbol{\beta}\mathbf{b}^T\mathbf{b}$$

to get ($\boldsymbol{\beta} =$ what? – guess at a solution); any multiple will do. Then

$$\boldsymbol{\alpha} = \mathbf{S}^{-1/2}\boldsymbol{\beta} = \mathbf{S}^{-1}\left(\bar{\mathbf{x}} - \bar{\mathbf{y}}\right)$$

and we classify as A if

$$|w - \bar{u}| = \left|\boldsymbol{\alpha}^T(\mathbf{z} - \bar{\mathbf{x}})\right| < \left|\boldsymbol{\alpha}^T(\mathbf{z} - \bar{\mathbf{y}})\right| = |w - \bar{v}|.$$

Equivalently (difference of squares, etc.),

$$(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T \mathbf{S}^{-1}\left(\mathbf{z} - \frac{\bar{\mathbf{x}} + \bar{\mathbf{y}}}{2}\right) > 0,$$

which is (20.3), i.e. Bayes' rule, if the priors and costs are equal.

– Note that $D^2 = \|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|^2_{\mathbf{S}}$ is a constant multiple of Hotelling's T^2, used to test that the two means are equal. Thus if it is not too large, we should not expect the method to be useful.

– We might also project using a second eigenvector.

- Given J populations $\pi_1, ..., \pi_J$ with means μ_j and common covariance $\mathbf{\Sigma}$, and a r.vec. $\mathbf{x}$ arising from one of them, consider the problem of finding linear functions $Y = \mathbf{a}^T\mathbf{x}$ which best separate the populations. We have

$$\begin{aligned} E\left[Y|\pi_j\right] &= \mathbf{a}^T\mu_j, \\ \text{cov}\left[Y|\pi_j\right] &= \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}. \end{aligned}$$

Assuming priors $\left\{p_j\right\}$, the unconditional mean is

$$\sum p_j E\left[Y|\pi_j\right] = \sum_{j=1}^{J} p_j \mathbf{a}^T\mu_j = \mathbf{a}^T\bar{\mu},$$

for

$$\bar{\mu} = \sum_{j=1}^{J} p_j\mu_j.$$

To best separate the populations we should choose $\mathbf{a}$ so as to maximize $\sum_{j=1}^{J} p_j\left(E\left[Y|\pi_j\right] - E\left[Y\right]\right)^2$, relative to var$[Y]$. Equivalently, we maximize

$$\frac{\sum_{j=1}^{J} p_j\left(\mathbf{a}^T\left(\mu_j - \bar{\mu}\right)\right)^2}{\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}} = \frac{\mathbf{a}^T\mathbf{B}_{\boldsymbol{\mu}}\mathbf{a}}{\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}} \overset{def}{=} D^2\left(\mathbf{a}\right), \tag{22.2}$$

where

$$\mathbf{B}_\mu = \sum_{j=1}^{J} p_j \left(\mu_j - \bar{\mu}\right)\left(\mu_j - \bar{\mu}\right)^T.$$

Since

$$\begin{aligned} \max \frac{\mathbf{a}^T \mathbf{B}_\mu \mathbf{a}}{\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}} &= \max_{\|\mathbf{e}\|=1} \mathbf{e}^T \boldsymbol{\Sigma}^{-1/2} \mathbf{B}_\mu \boldsymbol{\Sigma}^{-1/2} \mathbf{e} \\ &= ch_{\max} \boldsymbol{\Sigma}^{-1/2} \mathbf{B}_\mu \boldsymbol{\Sigma}^{-1/2}, \end{aligned}$$

the maximizing $\mathbf{a}$ is (any multiple of)

$$\mathbf{a}_1 = \boldsymbol{\Sigma}^{-1/2} \mathbf{e}_1,$$

where $\mathbf{e}_1$ is the eigenvector corresponding to the maximum eigenvalue of $\boldsymbol{\Sigma}^{-1/2} \mathbf{B}_\mu \boldsymbol{\Sigma}^{-1/2}$. Continuing, there are

$$s \le \min\left(rk\left(\mathbf{B}_\mu\right), rk\left(\boldsymbol{\Sigma}\right)\right) \overset{why?}{\le} \min\left(J-1, p\right)$$

non-zero eigenvalues

$$\lambda_1 \ge \cdots \ge \lambda_s > 0 = \lambda_{s+1} = \cdots = \lambda_p,$$

and the corresponding r.v.s $Y_k = \mathbf{a}_k^T \mathbf{x}$, where $\mathbf{a}_k = \boldsymbol{\Sigma}^{-1/2} \mathbf{e}_k$, are the first, second, ... *population discriminants*.

- Let $\mathbf{A}_{p\times p}$ have the $\{\mathbf{a}_k\}$ as its columns, i.e. $\mathbf{A} = \mathbf{\Sigma}^{-1/2}\mathbf{E}$. Since $\mathbf{A}^T\mathbf{B}_\mu\mathbf{A} = diag\left(\lambda_1, ...\lambda_s, 0, ..., 0\right)$, for $k > s$ we have
$$0 = \mathbf{a}_k^T\mathbf{B}_\mu\mathbf{a}_k = \sum_{j=1}^{J} p_j \left(\mathbf{a}_k^T\left(\mu_j - \bar{\mu}\right)\right)^2,$$
so that $\mathbf{a}_k^T\left(\mu_j - \bar{\mu}\right) = 0$ for all $j = 1, ..., J$. Thus the $\mathbf{a}_k$ for $k > s$ do not contribute to the classification – the means $E\left[Y_k|\pi_j\right] = \mathbf{a}_k^T\mu_j$ of these population discriminants are constant.

- If $\mathbf{x}$ arises from π_j then the vector $\mathbf{l}^{(p)}\left(\mathbf{x}\right) = \mathbf{A}^T\mathbf{x}$ of population discriminants has mean
$$E\left[\mathbf{l}^{(p)}\left(\mathbf{x}\right)|\pi_j\right] = \mathbf{A}^T\mu_j = \mathbf{l}^{(p)}\left(\mu_j\right)$$
and $\text{cov}\left[\mathbf{l}^{(p)}\left(\mathbf{x}\right)\right] = \mathbf{A}^T\mathbf{\Sigma}\mathbf{A} = \mathbf{I}_p$, and so we classify $\mathbf{x}$ into population
$$j^* = \arg\min_j \left\|\mathbf{l}^{(p)}\left(\mathbf{x}\right) - \mathbf{l}^{(p)}\left(\mu_j\right)\right\|^2 = \arg\min_j \left\|\mathbf{x} - \mu_j\right\|_{\mathbf{\Sigma}}^2.$$
This – ‘Fisher’s method’ – is again Bayes’ rule, if all populations are Normal, and if the priors are equal – but normality is not required in the derivation of Fisher’s method.

- The general 'Fisher's classification procedure' uses only the first $r \leq s$ of the discriminants. The rule is to classify $\mathbf{x}$ into population
$$j^* = \arg\min_j \left\| \mathbf{l}^{(r)}(\mathbf{x}) - \mathbf{l}^{(r)}\left(\mu_j\right) \right\|^2,$$
where $\mathbf{l}^{(r)}(\mathbf{x}) = \left[\mathbf{A}^{(r)}\right]^T \mathbf{x}$ (the first r columns of $\mathbf{A}$, hence the first r discriminants).

- Why the *first* r? Consider the population discriminants and the measure $D^2(\mathbf{a})$ at (22.2). This is maximized by $\mathbf{a}_1$, i.e. by using $Y_1 = \mathbf{a}_1^T \mathbf{x}$, with
$$\max D^2(\mathbf{a}) = D^2(\mathbf{a}_1) = \lambda_1.$$
Now we try to separate, as best we can, using information not contained in Y_1 – i.e. by a linear combination $\mathbf{a}^T\mathbf{x}$ uncorrelated with Y_1:
$$\mathbf{a} = \arg\max_{\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}_1=0} D^2(\mathbf{a}).$$
With $\mathbf{e} = \boldsymbol{\Sigma}^{1/2}\mathbf{a}$ as before this becomes
$$\mathbf{e} = \arg\max_{\mathbf{e}^T\mathbf{e}_1=0} \mathbf{e}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{B}_\mu\boldsymbol{\Sigma}^{-1/2}\mathbf{e} = \mathbf{e}_2,$$

with

$$\mathbf{a} = \mathbf{\Sigma}^{-1/2}\mathbf{e}_2 = \mathbf{a}_2,$$
$$D^2(\mathbf{a}_2) = \lambda_2,$$

and $Y_2 = \mathbf{a}_2^T\mathbf{x}$, the second population discriminant. Continue (as in Lecture 14 – pc's) ... ; conclude that at any point the separation is best accomplished by using the first discriminant not yet employed.

- A plausible guide to the choice of r is

  $$\frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{s}\lambda_k} = \frac{\sum_{k=1}^{r}\lambda_k}{tr\left[\mathbf{\Sigma}^{-1}\mathbf{B}_{\boldsymbol{\mu}}\right]};$$

  in practice these are of course replaced by their sample equivalents.

- Given training samples from each population we first compute sample means $\bar{\mathbf{x}}_j$, the overall mean $\bar{\mathbf{x}} = \sum_{j=1}^{J} n_j \bar{\mathbf{x}}_j / n$ and

$$\begin{aligned} \mathbf{B} &= \sum_{j=1}^{J} n_j \left(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}\right)^T, \\ \mathbf{W} &= \sum_{j=1}^{J} \sum_{k=1}^{n_j} \left(\mathbf{x}_{kj} - \bar{\mathbf{x}}_j\right)\left(\mathbf{x}_{kj} - \bar{\mathbf{x}}_j\right)^T \\ &= (n - J)\, \mathbf{S}_{pooled}. \end{aligned}$$

  Then determine the eigenvectors $\{\hat{\mathbf{e}}_k\}$ of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$; put $\hat{\mathbf{a}}_k = \mathbf{W}^{-1/2}\hat{\mathbf{e}}_k$. Let $\hat{\mathbf{l}}^{(r)}(\mathbf{x}) = \left[\hat{\mathbf{A}}^{(r)}\right]^T \mathbf{x}$, where

$$\hat{\mathbf{A}}^{(r)} = \left(\hat{\mathbf{a}}_1 \vdots \cdots \vdots \hat{\mathbf{a}}_r\right);$$

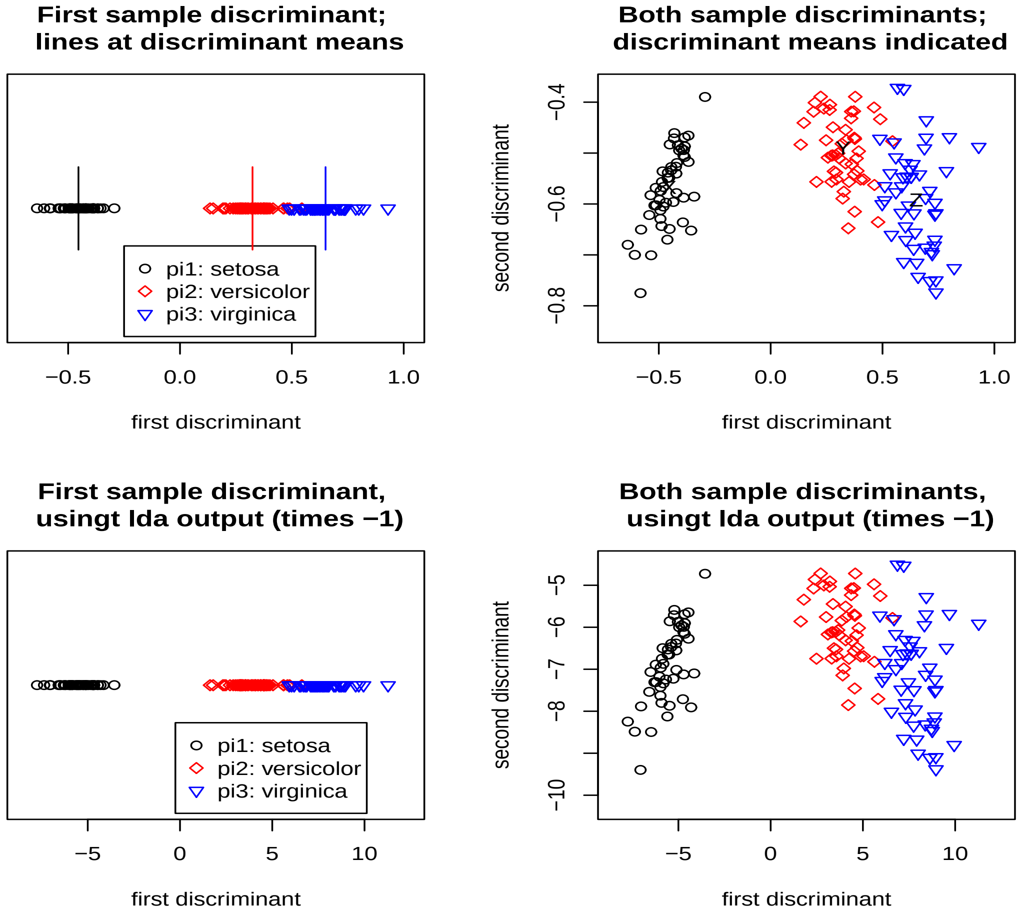
  classify $\mathbf{x}$ into population

$$j^* = \arg\min_j \left\| \hat{\mathbf{l}}^{(r)}(\mathbf{x}) - \hat{\mathbf{l}}^{(r)}\left(\bar{\mathbf{x}}_j\right) \right\|^2.$$

  - Particularly when $r = 1, 2$ plots of the *sample discriminants* – the elements of $\hat{\mathbf{l}}^{(r)}(\mathbf{x}_i)$ for $i = 1, ..., n$ – can be quite informative.

- Example: see web site. The `lda()` function in R returns the sample discriminants as the object `$scaling`. They can be on a different scale than those calculated exactly as described above, but otherwise agree with them if the prior probabilities are equal.

```
Sample discriminants calculated above are
      [,1]  [,2]
[1,] -0.07  0.00
[2,] -0.13 -0.18
[3,]  0.18  0.08
[4,]  0.23 -0.23
Sample discriminants returned by lda are
             LD1   LD2
sep.length  0.83  0.02
sep.width   1.53  2.16
pet.length -2.20 -0.93
pet.width  -2.81  2.84
```



First sample discriminant;
lines at discriminant means

Both sample discriminants;
discriminant means indicated

First sample discriminant,
usingt lda output (times −1)

Both sample discriminants,
usingt lda output (times −1)

23. Clustering: measures and methods

- **Purpose**. Given p-dimensional observations, group these into g significantly distinct groups, within which they are homogeneous (similar). In distinction to the situation in discrimination, here both the number of groups and the reasons for putting an observation into one group rather than another are unknown. The method is largely exploratory, with the intention of following up with a more detailed analysis of the groups.

- An immediate problem is to define the notion of similarity. Major methods include 'distances' for continuous variables and 'similarity coefficients' for discrete variables.

- **Distances between pairs of items**. Common measures are $d(\mathbf{x},\mathbf{y}) =$

  $$\|\mathbf{x}-\mathbf{y}\| = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \text{(Euclidean distance)},$$
  $$\sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})} \text{ (Mahalanobis distance)},$$
  $$\left(\sum |x_i - y_i|^m\right)^{1/m} \text{ (Minkowski distance; } L^m\text{)},$$
  $$\sum |x_i - y_i| \text{ (City-block distance; } L^1\text{)}.$$

  These are true distances:

  $$\begin{aligned} \text{(i) } d(\mathbf{x},\mathbf{y}) &\geq 0\text{, equality iff } \mathbf{x}=\mathbf{y}, \\ \text{(ii) } d(\mathbf{x},\mathbf{y}) &= d(\mathbf{y},\mathbf{x}), \\ \text{(iii) } d(\mathbf{x},\mathbf{z}) &\leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}). \end{aligned}$$

- **Similarity coefficients for pairs of items**. Suppose items $\mathbf{x}$ and $\mathbf{y}$ are represented by binary elements, i.e. $\{0, 1\}$, with a '1 ' indicating the presence of a certain characteristic. Record the numbers of matches and mismatches:

		y		
		1	0	
x	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	p

If, e.g., the variable is recording a rare attribute then a match might be much more significant than a mismatch, leading to similarity coefficients such as:

$\dfrac{a+d}{p}$ (Equal weights for 1-1 and 0-0; $= 1 - \dfrac{\|\mathbf{x}-\mathbf{y}\|^2}{p}$),
$\dfrac{2\,(a+d)}{2\,(a+d)+b+c}$ (Matches more heavily weighted than mismatches),
$\dfrac{a}{p}$ (Only 1-1 matches count),
etc.; see Table 12.1 in the text.

If $d(\mathbf{x}, \mathbf{y})$, with $d_{ij} = d(\mathbf{x}_i, \mathbf{y}_j)$, is a distance, then

$$s_{ij} = \frac{1}{1 + d_{ij}} \in [0, 1]$$

defines a measure of ‘similarity’ between $\mathbf{x}_i$ and $\mathbf{y}_j$ (i.e. $s_{ij} \in [-1, 1]$ or $[0, 1]$, $s_{ij} = 1$ iff $\mathbf{x}_i = \mathbf{y}_j$). If the ‘similarity matrix’, with elements s_{ij}, is nnd, then (assigned)

$$d(\mathbf{x}_i, \mathbf{y}_j) = d_{ij} = \sqrt{2\left(1 - s_{ij}\right)} \tag{23.1}$$

is a distance, i.e. satisfies (i)-(iii) above.

- Suppose there are n observations $\mathbf{x}_1, ..., \mathbf{x}_n$, each of which is a binary *variable*: $\mathbf{x}_k = \left(U_{k1}, ..., U_{kp}\right)$, with $U_{ki} \in \{0, 1\}$ and frequencies as in the following table; for instance $c = \#\left\{k | \left(U_{ki}, U_{kj}\right) = (0, 1)\right\}$.

		variable j		
		1	0	
variable	1	a	b	$a+b$
i	0	c	d	$c+d$
		$a+c$	$b+d$	n

Then (assigned) the sample correlation coefficient between U_i and U_j is

$$r_{ij} = \frac{ad - bc}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}}, \qquad (23.2)$$

and $\left\{r_{ij}\right\}$ defines a similarity measure $\in [-1, 1]$ between the variables (with variables U_i and U_j being called equal if $b = c = 0$ in the table).

- All this is very subjective; see Example 12.6, discussed below, for a different approach tailored to a particular application.

- **Hierarchical clustering**. Here clusters are formed sequentially, with the number of clusters decreasing as clusters are merged with other similar clusters (*agglomerative* hierarchical methods) or split into less homogeneous groups (*divisive* methods). Only the former is discussed here.

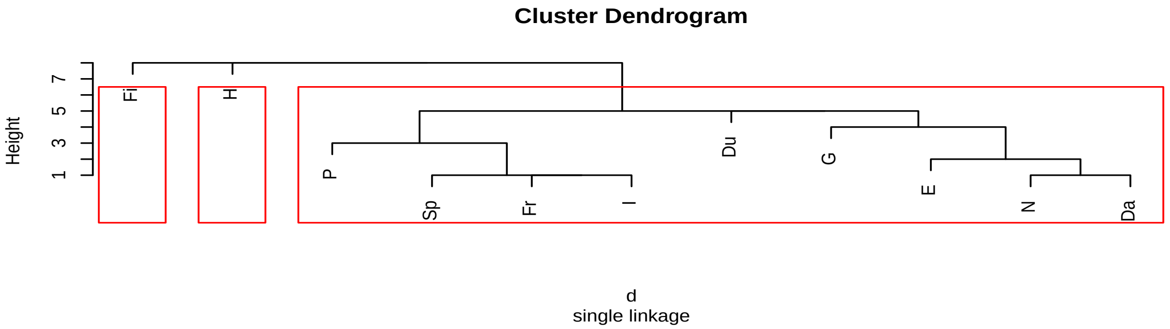
- **Agglomerative clustering**. Initially every case is a cluster of size one; given clusters $C_1, ..., C_k$, the 'most similar' pair of clusters is merged. Continue. This requires a 'linkage' allowing us to extend the distance (or similarity) between cases or variables to that between clusters. Possibilities are $d\left(C, D\right) =$

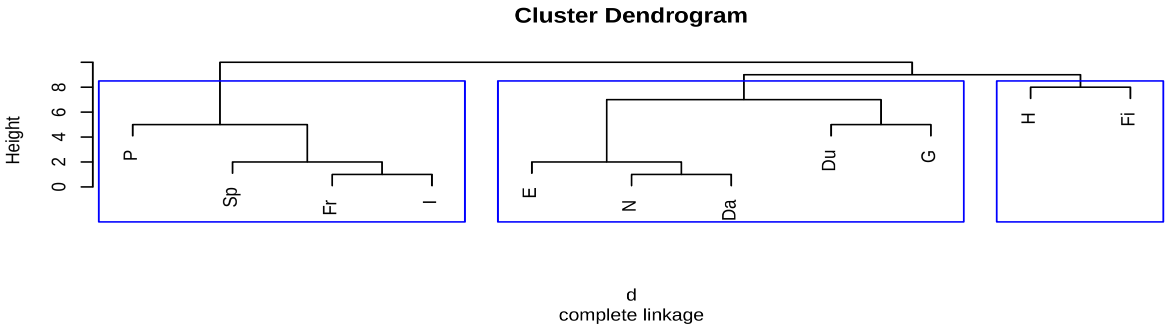
$$\begin{array}{l}
\min_{\mathbf{x}\in C, \mathbf{y}\in D} d(\mathbf{x}, \mathbf{y}) \text{ ('single' linkage)} \\
\max_{\mathbf{x}\in C, \mathbf{y}\in D} d(\mathbf{x}, \mathbf{y}) \text{ ('complete' linkage)} \\
aver_{\mathbf{x}\in C, \mathbf{y}\in D} d(\mathbf{x}, \mathbf{y}) \\
median_{\mathbf{x}\in C, \mathbf{y}\in D} d(\mathbf{x}, \mathbf{y}) \\
d(\bar{\mathbf{x}}_C, \bar{\mathbf{y}}_D) \text{ ('centroid' linkage).}
\end{array}$$

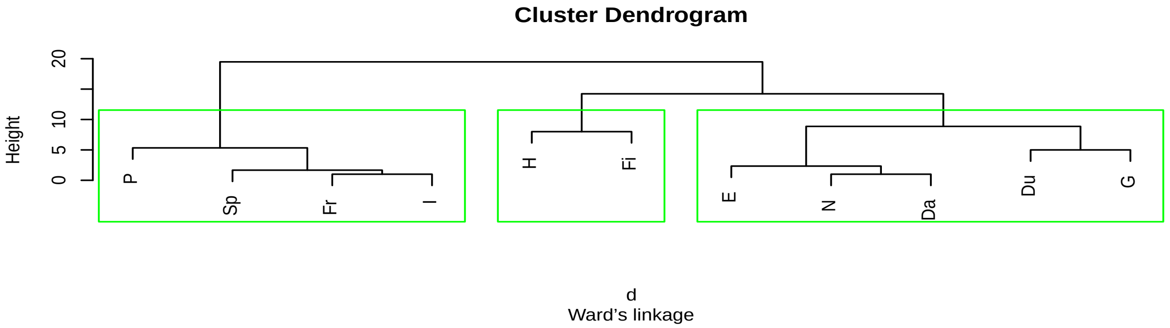
- Example 12.6 – The numbers 1 - 10 are written in each of 11 languages, and the similarity measure $s(\mathbf{x}, \mathbf{y})$ is the number of times that the first letter of the numeral 'i' in language 'X' is the same as that in the language 'Y' ($i = 1, ..., 10$). For instance $s(Eng, Fr) = 4$, arising from $\{$(three,trois),(six,six),(seven,sept),(nine,neuf)$\}$. Then $d(\mathbf{x}, \mathbf{y}) = 10 - s(\mathbf{x}, \mathbf{y}) \in [0, ..., 10]$. The R function

  ```
     fit   =   hclust(d, method = "single")
  groups   =    cutree(fit, k=3)
  ```

  (here `d` is a 'distance' object, there are other linkages possible) followed by `plot(fit)` gives the following 'dendograms', with three clusters shown in each method. Code on web site.


Cluster Dendrogram
Height
1
3
5
7
Fi
H
P
Sp
Fr
I
Du
G
E
N
Da
d
single linkage


Cluster Dendrogram
Height
0
2
4
6
8
P
Sp
Fr
I
E
N
Da
Du
G
H
Fi
d
complete linkage


Cluster Dendrogram
Height
0
5
10
20
P
Sp
Fr
I
H
Fi
E
N
Da
Du
G
d
Ward’s linkage

- **Ward's method** – an agglomerative method aimed at finding elliptical clusters (on the assumption that the data are approximately multivariate normal). If there are currently K clusters, then let ESS_k be the SS (around the mean) in the k^{th} cluster:

  $$ESS_k = \sum_{\mathbf{x} \in C_k} \|\mathbf{x} - \bar{\mathbf{x}}_k\|^2 ,$$

  and $ESS = \sum_k ESS_k$. At each stage, the two clusters to be joined are those that result in the smallest increase in ESS. Initially $ESS = 0$ (each observation is its own cluster centroid), eventually

  $$ESS = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = tr\left[(n-1)\,\mathbf{S}\right]$$

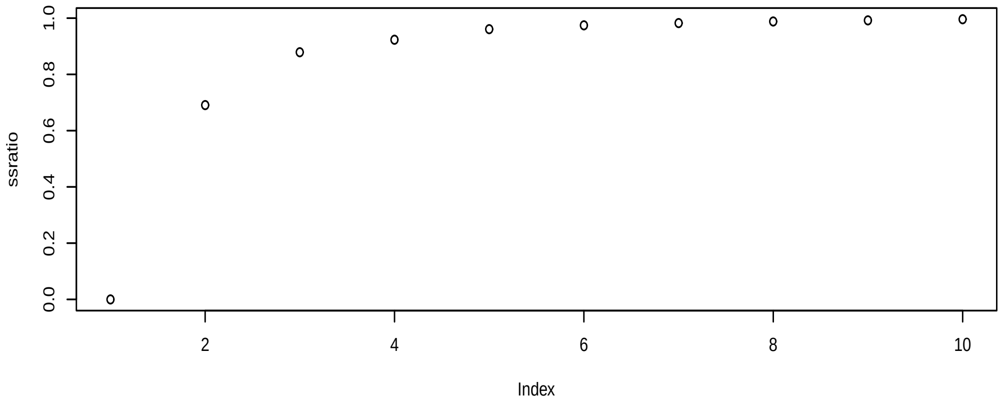
  (there is only one cluster). The user specifies the desired number of clusters, hence the stopping point. This is an optional method in `hclust`.

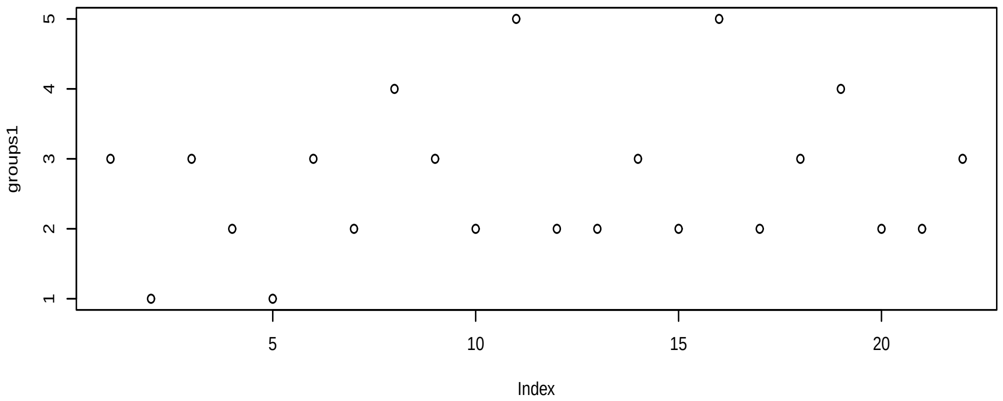
- **Nonhierarchical methods**. Only the ‘K-means’ method is discussed here.
  - Initialization: Start with a partition of the data into K clusters, or specify K initial centroids.
  - Iterative step: Assign each object to the cluster whose centroid is closest to the object (as in classification). This results in K new clusters.
  - Recalculate the centroids and repeat the iterative step until there are no more changes. (To preserve the value of K one might split large clusters into two, if one cluster becomes empty.)

- The population version of the underlying principle is that, for a p-dimensional random vector $\mathbf{x}$, we seek points $\xi_1, ..., \xi_K \in \mathbb{R}^p$ to minimize
$$E\left[\min_k \|\mathbf{x} - \xi_k\|^2\right].$$
The sample version is to minimize
$$aver_i\left[\min_k \|\mathbf{x}_i - \xi_k\|^2\right].$$

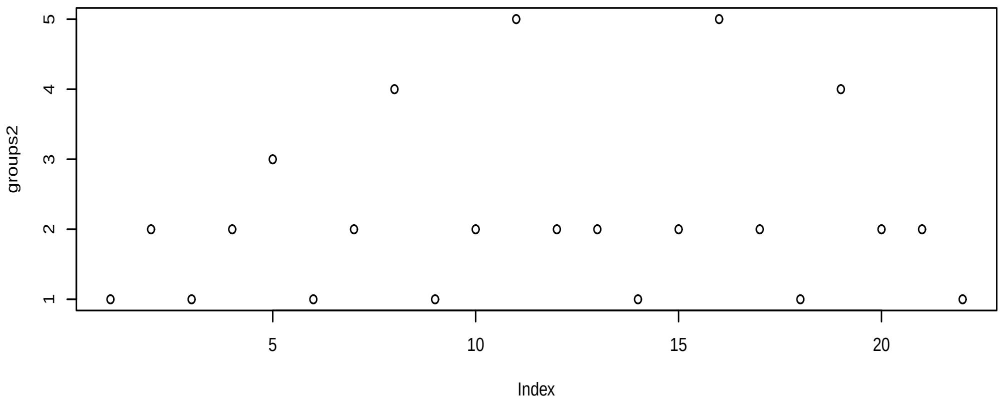
- Choice of K is adaptive – plotting? maximizing some function of $\mathbf{W}^{-1}\mathbf{B}$ (the 'within' and 'between' SSCP matrices)?

- Example 12.12 – Data on 22 public utilities in Table 12.4; $p = 8$ measurements on each – sales, % nuclear, etc. Use `fit = kmeans(X, k, nstart)`. (`nstart` $= \#$ of randomly chosen starting configurations.) Plot $|\mathbf{B}| / |\mathbf{B} + \mathbf{W}|$ (a measure of the overall difference between the ξ_k; $\mathbf{B}$ and $\mathbf{W}$ are returned by `fit`) vs. k; choose $k = 5$. Results are somewhat similar to using hierarchical clustering with centroid linkage (except for the group labels). Code on web site.





k−means



hierarchical, centroid linkage

24. Model based clustering; multidimensional scaling

- **Clustering by likelihood methods**. Assume that the density of the population is a mixture
$$\sum_{k=1}^{K} p_k f\left(\mathbf{x}|\theta_k\right)$$
with (at first) K known. For instance $f\left(\mathbf{x}|\theta_k\right)$ could be the $N_p\left(\mu_k, \Sigma_k\right)$ density. Estimate the $\left\{p_k, \theta_k\right\}_{k=1}^{K}$ by maximum likelihood. Then K is chosen to maximize the penalized likelihood
$$\begin{aligned} BIC &= 2\log L - 2\log\left\{N\left(\frac{K\left(p+1\right)\left(p+2\right)}{2}-1\right)\right\} \\ &\approx 2\log\left(L/K\right) + c_{N,p}. \end{aligned}$$

- To fit by ML: For any assignment of groups, define $\gamma_i = k$ if $\mathbf{x}_i$ is from the k^{th} group and let $\gamma = \left(\gamma_1, ..., \gamma_n\right)$; then the likelihood is
$$L\left(\gamma, \theta_1, ..., \theta_K\right) = \prod_{k=1}^{K} \prod_{\{i|\gamma_i=k\}} f\left(\mathbf{x}_i|\theta_k\right) = \prod_{i=1}^{N} f\left(\mathbf{x}_i|\theta_{\gamma_i}\right).$$

Let $\hat{\theta}(\gamma)$ be the mle for θ given γ, then $\hat{\gamma}$ maximizes $L\left(\gamma, \hat{\theta}(\gamma)\right)$. For instance in Normal models $\theta_k = (\mu_k, \mathbf{\Sigma}_k)$, $\hat{\mu}_k$ and $\hat{\mathbf{\Sigma}}_k$ are the average and covariance of the observations with $\gamma_i = k$ and

$$\log L\left(\gamma, \hat{\theta}(\gamma)\right) = const - \frac{1}{2}\sum_{k=1}^{K} n_k(\gamma) \log |\mathbf{S}_k(\gamma)| \,.$$

Finally, $\hat{p}_k = n_k(\hat{\gamma})/n$. One might impose restrictions such as $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K$.

- Can be done iteratively: Form groups (i.e., initialize γ) , estimate $\{p_k, \theta_k\}$, form groups again, iterate to convergence. The groups are formed each time by assigning an observation $\mathbf{x}$ to the group k for which

  $$\hat{p}(k|\mathbf{x}) = \hat{p}_k f\left(\mathbf{x}|\hat{\theta}_k\right) \Big/ \sum_{j=1}^{K} \hat{p}_j f\left(\mathbf{x}|\hat{\theta}_j\right)$$

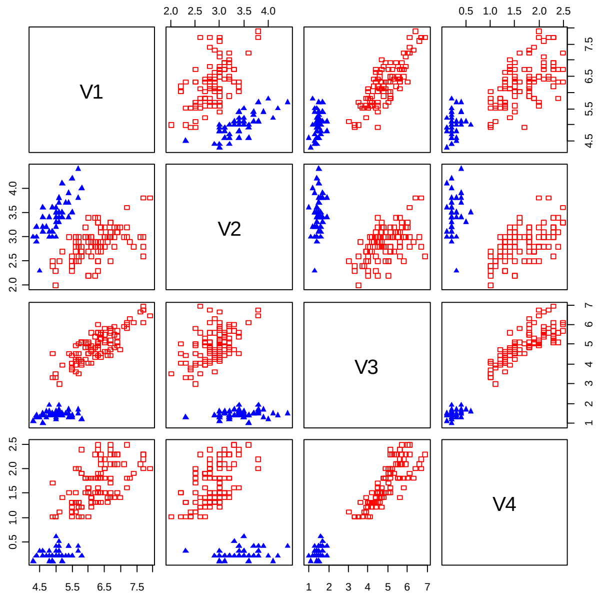
  is largest. [Typo in text here.]

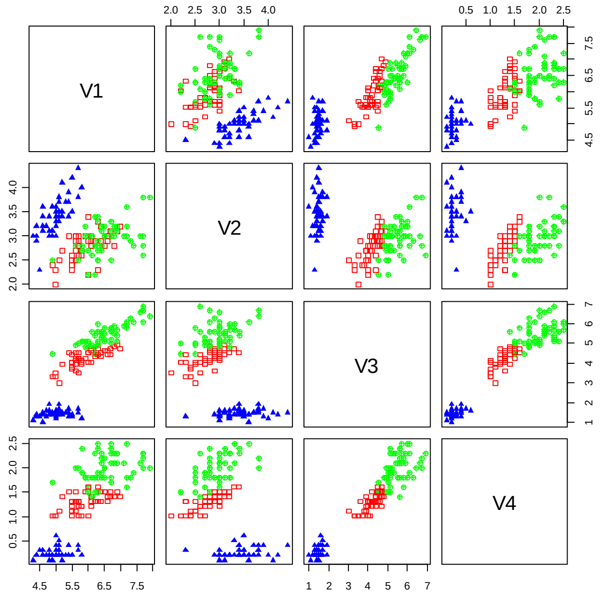
- Example 12.13 – Iris data. Code on web site; algorithm described in a paper also on web site.

Default method chooses 2 groups (rather than the known value $K = 3$); forcing $K = 3$ gives familiar plots.

```
> fit3$parameters
(V1 and V3 are the petal and sepal lengths)
$pro
[1] 0.333332 0.666668
$mean
        [,1]     [,2]
V1 5.0060021 6.261996
V2 3.4280046 2.871999
V3 1.4620006 4.905993
V4 0.2459998 1.675997

> fit4$parameters
$pro
[1] 0.3333333 0.3003844 0.3662823
$mean
    [,1]     [,2]     [,3]
V1 5.006 5.914879 6.546670
V2 3.428 2.777504 2.949495
V3 1.462 4.203758 5.481901
V4 0.246 1.298819 1.985322
```


V1
V2
V3
V4
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7


V1
V2
V3
V4
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
1 2 3 4 5 6 7

- **Multidimensional scaling**. Given n points in $\mathbb{R}^p$, and similarities $\left\{s_{ij}\right\}_{i<j}$, find points $\mathbf{y}_1, ..., \mathbf{y}_n \in \mathbb{R}^q$ (with q small) so that the distances $\left\{d_{ij}^{(q)}\right\}_{i<j}$ are related inversely to the $\left\{s_{ij}\right\}_{i<j}$ (i.e. ideally $d_{ij}^{(q)} < d_{i'j'}^{(q)}$ iff $s_{ij} > s_{i'j'}$). The idea is to find a low-dimensional representation of the data which mimics the original distances.

- Let $\left\{\hat{d}_{ij}^{(q)}\right\}_{i<j}$ be numbers – <u>not necessarily distances</u> – known to be exactly inversely related to the $\left\{s_{ij}\right\}_{i<j}$ (such as $\hat{d}_{ij}^{(q)} = \sqrt{2\left(1 - s_{ij}\right)}$?). There may not be q-dimensional points with these distances. If $\left\{d_{ij}^{(q)}\right\}_{i<j}$ are distances which *are* attained by q-dimensional points, their quality is measured by

$$Stress(q) = \left\{\frac{\sum_{i<j}\left(d_{ij}^{(q)} - \hat{d}_{ij}^{(q)}\right)^2}{\sum_{i<j}\left(d_{ij}^{(q)}\right)^2}\right\}^{1/2}.$$

An alternate measure $SStress(q)$ replaces each distance by its square, in this expression.

- Algorithm:

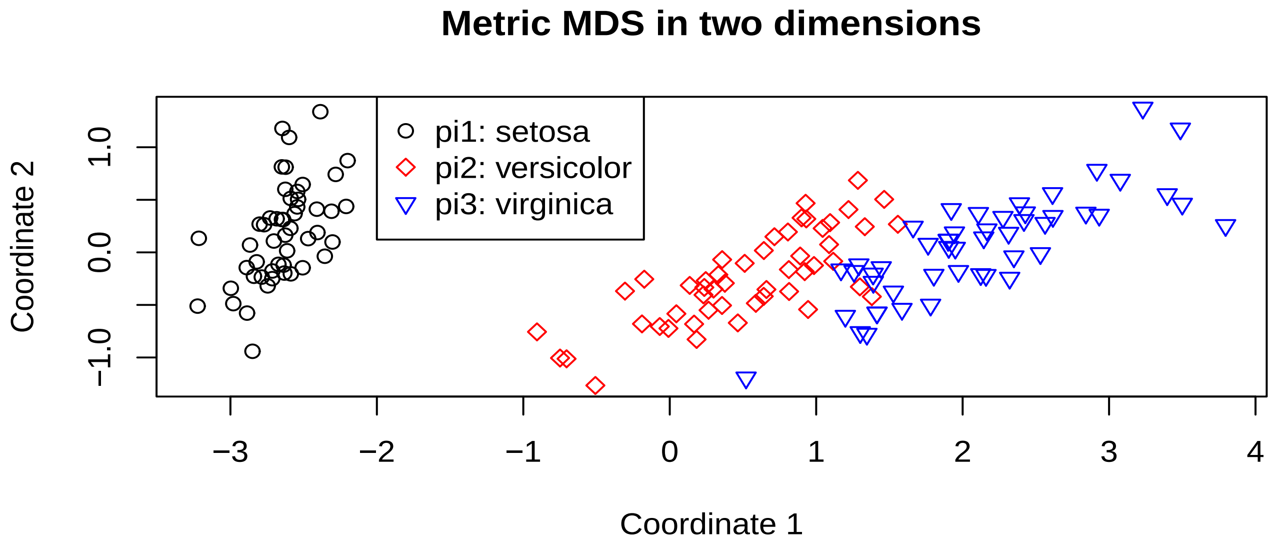
  - Take a trial configuration $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ of q-dimensional points. Compute the $\left\{d_{ij}^{(q)}\right\}$ and find numbers $\left\{\hat{d}_{ij}^{(q)}\right\}$ which minimize (the numerator of) $Stress(q)$, subject to the condition that they be exactly inversely related to the $\left\{s_{ij}\right\}$ (which are computed from the current $\left\{d_{ij}^{(q)}\right\}$) – isotonic regression?

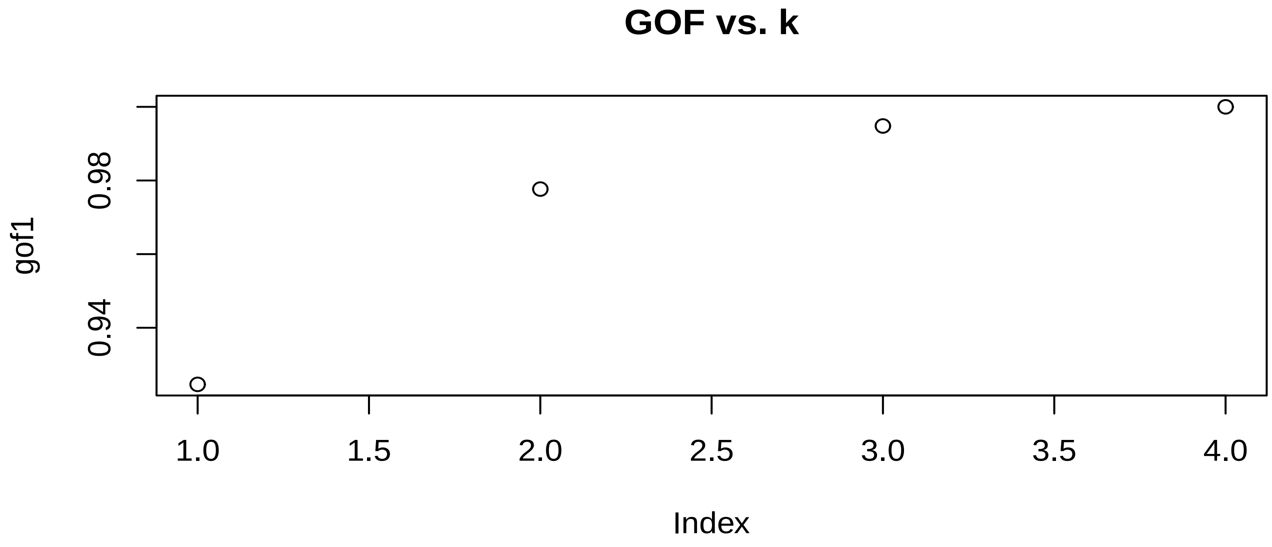
  - Using these new numbers $\left\{\hat{d}_{ij}^{(q)}\right\}$, find a new configuration $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ with an improved value of $Stress(q)$. (The stress is a function of the nq elements of the $\{\mathbf{y}_i\}$, to be minimized numerically.)

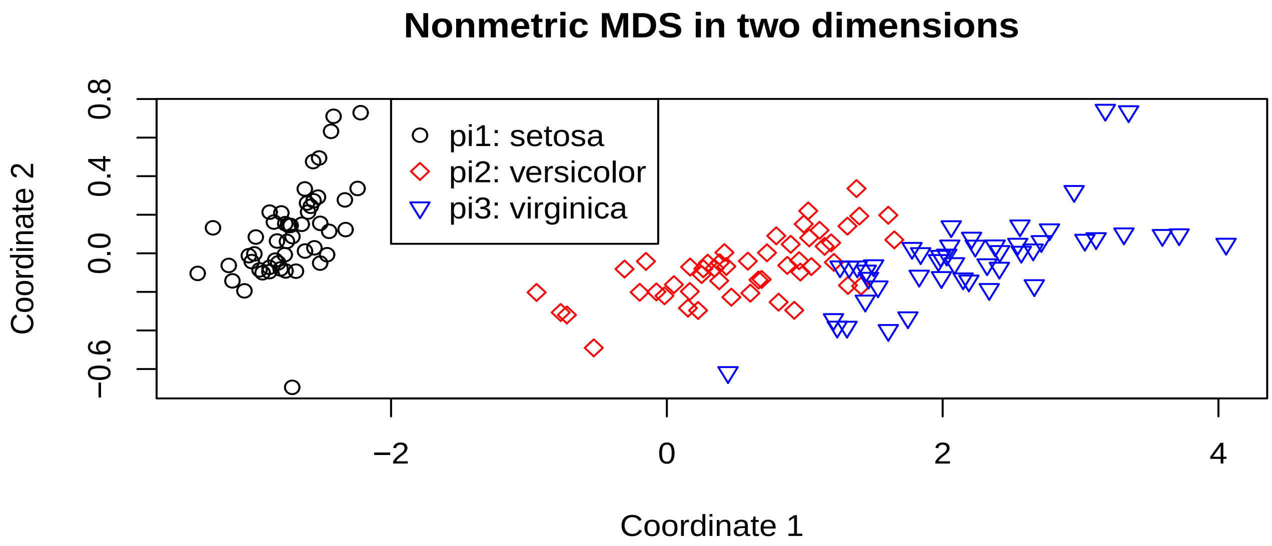
  - Iterate to convergence.

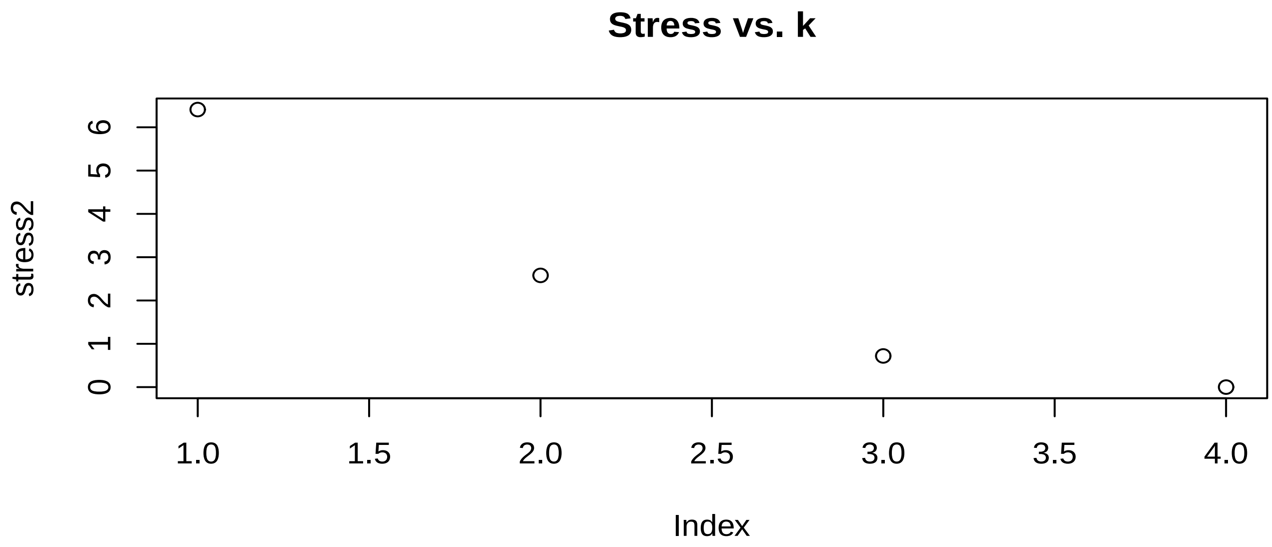
- This describes ‘metric multidimensional scaling’; if, in the computation of $\left\{s_{ij}\right\}$, the ordered similarities are replaced by their ranks then the procedure is ‘nonmetric multidimensional scaling’.

- Example – Iris data again. The R function `cmdscale(d,eig=TRUE, k)`, where `d` is a distance matrix, does metric mds; load the MASS library and use `isoMDS(d, y = cmdscale(d, k), k)` to do nonmetric mds. See web site. Both plots use L^1 distances; Euclidean distances give similar results.


Metric MDS in two dimensions
pi1: setosa
pi2: versicolor
pi3: virginica
Coordinate 2
−1.0
0.0
1.0
−3
−2
−1
0
1
2
3
4
Coordinate 1


GOF vs. k
gof1
0.94
0.98
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index


Nonmetric MDS in two dimensions
○ pi1: setosa
◇ pi2: versicolor
▽ pi3: virginica
Coordinate 2
−0.6
0.0
0.4
0.8
−2
0
2
4
Coordinate 1

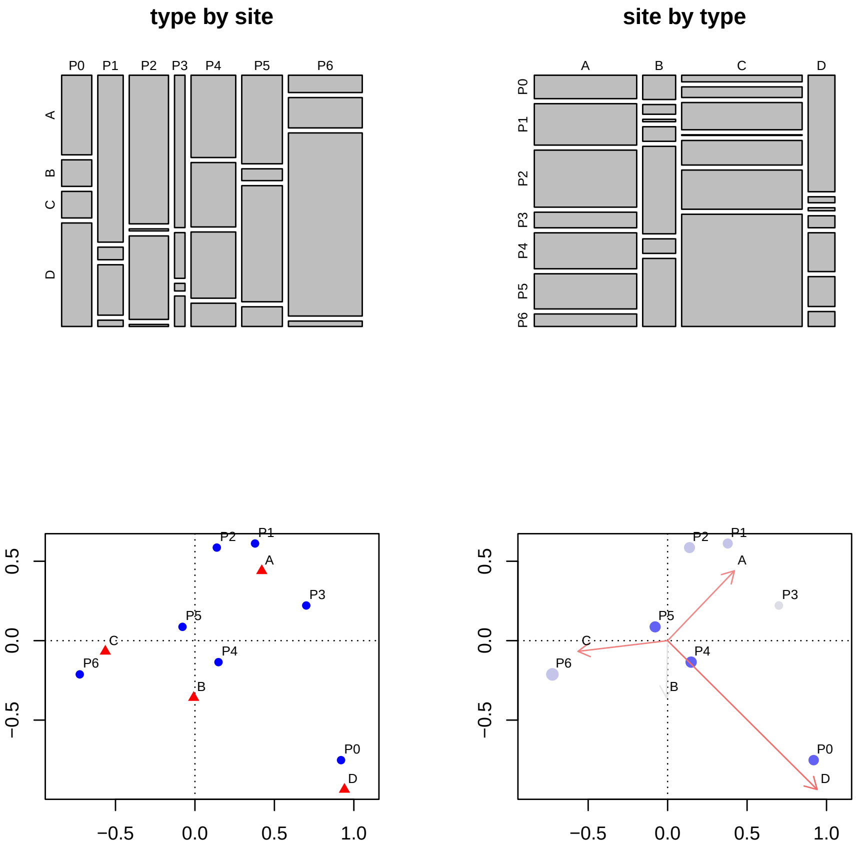

Stress vs. k
stress2
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Index

25. Correspondence analysis

- Suppose we are given a 2-way table of counts (a contingency table) with I rows and J columns. Example 12.17 (code on web site) – pottery types A – D are found at archeological sites P0 – P6, with the following frequencies:

	A	B	C	D
P0	30	10	10	39
P1	53	4	16	2
P2	73	1	41	1
P3	20	6	1	4
P4	46	36	37	13
P5	45	6	59	10
P6	16	28	169	5

Plots of the profiles (= conditional distribution of type given site, etc.) suggest that the types are not distributed over the sites in the same way. This has archeological significance.



Top: Sites P1 and P2 are similar; P0 and P6 are very different. Site P0 and type D seems to be strongly associated, as do P6 and C.

Bottom: Gives similar information in a more obvious fashion. Right plot uses more options; see help(plot.ca).

We seek a measure of the quality of a k-dimensional representation of the (‘centred’, i.e. beyond independence) associations. For instance – how much information is lost in representing the data by a two-dimensional plot – rather than three, where (we will see) it could be represented exactly? This is obtained from the ‘inertias’:

```
                  1        2        3
Inertia     0.283588 0.170107 0.058786
Percentage    55.34%   33.19%   11.47%
```

Interpretation: We might ‘explain’ the table merely through the row and column averages: $x_{ij} \overset{?}{=} n\bar{x}_{i\cdot}\bar{x}_{\cdot j}$, as would suffice if the attributes were independent (‘O_{ij} vs. E_{ij}’). This employs one of the $J = 4$ singular values; of the others the first – $\lambda_1^2 = .2836$ – represents 55.34% of the ‘inertia’; we call λ_1^2 the inertia associated with the first dimension. Similarly the inertia associated with the second dimension is 33.19%; together these account for 88% of the total – little information is lost when the ‘best’ two-dimensional representation is used.

Algebraic development. Let $\mathbf{X}$ be the $I \times J$ table, with $I > J = rk(\mathbf{X})$. Put $\mathbf{P}_{I\times J} = \mathbf{X}/n$, so that $\sum_{i,j} p_{ij} = 1$. Let $\mathbf{r}_{I\times 1} = \mathbf{P}\mathbf{1}_J$ be the vector of row totals, and $\mathbf{c}_{J\times 1} = \mathbf{P}^T\mathbf{1}_I$ the vector of column totals. Define $\mathbf{D_r}$ and $\mathbf{D_c}$ to be the diagonal matrices with diagonal elements $\{r_i\}_{i=1}^I$ and $\left\{c_j\right\}_{j=1}^J$ respectively. Finally, define

$$\mathbf{B}_{I\times J} = \mathbf{D_r}^{-1/2}\mathbf{P}\mathbf{D_c}^{-1/2},$$

with elements

$$b_{ij} = \frac{p_{ij}}{\sqrt{r_i c_j}}.$$

If $\hat{\mathbf{X}}$ is another, approximating, $I \times J$ table with the same row and column totals as $\mathbf{X}$, and $\hat{\mathbf{P}}$, $\hat{\mathbf{B}}$ are formed as above, then the quality of the approximation can be measured by

$$\begin{aligned} \sum_{i,j} \frac{\left(p_{ij} - \hat{p}_{ij}\right)^2}{r_i c_j} &= \sum_{i,j}\left(b_{ij} - \hat{b}_{ij}\right)^2 \\ = \;& tr\left[\left(\mathbf{B} - \hat{\mathbf{B}}\right)^T\left(\mathbf{B} - \hat{\mathbf{B}}\right)\right] = \left\|\mathbf{B} - \hat{\mathbf{B}}\right\|^2 \\ (\; = \;& \chi^2/n \text{ if } \hat{p}_{ij} = r_i c_j). \end{aligned}$$

- **Result 1**: Let

$$\mathbf{B} = \mathbf{U}_{I\times I}\mathbf{D}_{I\times J}\mathbf{V}^T_{J\times J} = \sum_{j=1}^{J} \lambda_j \mathbf{u}_j \mathbf{v}_j^T$$

be the singular value decomposition, where

$$\mathbf{D} = \left(\begin{array}{c} \Lambda_{J\times J} \\ \mathbf{0}_{(I-J)\times J} \end{array} \right)$$

and $\lambda_1^2 \geq \cdots \geq \lambda_J^2 > 0$ are the eigenvalues of $\mathbf{B}^T\mathbf{B}$. Among all matrices $\mathbf{C}_{I\times J}$ of rank at most $k < J$, $\|\mathbf{B} - \mathbf{C}\|^2$ is minimized by

$$\hat{\mathbf{B}} = \sum_{j=1}^{k} \lambda_j \mathbf{u}_j \mathbf{v}_j^T.$$

Proof: For any such $\mathbf{C}$, we have

$$\begin{aligned} \|\mathbf{B} - \mathbf{C}\|^2 &= \left\| \mathbf{D} - \mathbf{U}^T\mathbf{C}\mathbf{V} \right\|^2 \\ &= \left\| \mathbf{D} - \tilde{C} \right\|^2 \\ &= \left\| \left(\begin{array}{c} \Lambda \\ \mathbf{0} \end{array} \right) - \left(\begin{array}{c} \tilde{C}_1 \\ \tilde{C}_2 \end{array} \right) \right\|^2 \\ &= \left\| \Lambda - \tilde{C}_1 \right\|^2 + \left\| \tilde{C}_2 \right\|^2, \end{aligned}$$

in an obvious notation. This continues as

$$\|\mathbf{B}-\mathbf{C}\|^2 = \\ \sum_{j=1}^{J}\left(\lambda_j-\tilde{C}_{1,jj}\right)^2+\sum_{i\neq j=1}^{J}\tilde{C}_{1,ij}^2+\left\|\tilde{C}_2\right\|^2.$$

This is minimized by a matrix $\tilde{C}$ with the same structure as $\mathbf{D}$, with its diagonal elements chosen to minimize $\sum_{j=1}^{J}\left(\lambda_j-\tilde{C}_{1,jj}\right)^2$, subject to the restriction that at most k of them be nonzero. Since the λ_j are ordered from largest to smallest, this forces $\tilde{C}_{1,jj}=\lambda_j$ for $j=1,..,k$, and $=0$ for $j>k$. Then

$$\mathbf{C}=\mathbf{U}\tilde{\mathbf{C}}\mathbf{V}^T=\hat{\mathbf{B}}.$$

- **Result 2**: In the same notation as above, the largest singular value is $\lambda_1=1$, and the corresponding left and right eigenvectors are

$$\begin{aligned}\mathbf{u}_1 &= \mathbf{D}_\mathbf{r}^{1/2}\mathbf{1}_I=\mathbf{D}_\mathbf{r}^{-1/2}\mathbf{r},\\ \mathbf{v}_1 &= \mathbf{D}_\mathbf{c}^{1/2}\mathbf{1}_J=\mathbf{D}_\mathbf{c}^{-1/2}\mathbf{c}.\end{aligned}$$

Proof: First note that for these choices of $\mathbf{u_1}$ and $\mathbf{v_1}$,

$$\begin{aligned} \mathbf{B}^T\mathbf{B}\mathbf{v_1} &= \mathbf{v_1}, \\ \mathbf{B}\mathbf{B}^T\mathbf{u_1} &= \mathbf{u_1}. \end{aligned}$$

For instance

$$\begin{aligned} \mathbf{B}^T\mathbf{B}\mathbf{v_1} &= \mathbf{B}^T\left(\mathbf{D_r}^{-1/2}\mathbf{P}\mathbf{D_c}^{-1/2}\right)\mathbf{D_c}^{-1/2}\mathbf{c} \\ &= \mathbf{B}^T\mathbf{D_r}^{-1/2}\mathbf{P}\mathbf{1}_J \\ &= \mathbf{B}^T\mathbf{D_r}^{-1/2}\mathbf{r} \\ &= \left(\mathbf{D_c}^{-1/2}\mathbf{P}^T\mathbf{D_r}^{-1/2}\right)\mathbf{D_r}^{-1/2}\mathbf{r} \\ &= \mathbf{D_c}^{-1/2}\mathbf{P}^T\mathbf{1}_I \\ &= \mathbf{D_c}^{-1/2}\mathbf{c} \\ &= \mathbf{v_1}. \end{aligned}$$

Thus $\mathbf{u_1}$ and $\mathbf{v_1}$ are eigenvectors of $\mathbf{B}\mathbf{B}^T$ and $\mathbf{B}^T\mathbf{B}$ respectively, with eigenvalue 1 ($= \lambda_1^2$). It remains to show that this is the *largest* of the $\left\{\lambda_j^2\right\}_{j=1}^J$. Equivalently, for any $\mathbf{x}_{J\times 1}$,

$$\|\mathbf{B}\mathbf{x}\|^2 \leq \|\mathbf{x}\|^2 .$$

For this, first note that $\left\{q_{ij} = p_{ij}/r_i\right\}$ is a probability distribution on $\{1, ..., J\}$. Then by the Cauchy-Schwarz Inequality,

$$\begin{aligned}
\|\mathbf{Bx}\|^2 &= \sum_i \left\{ \sum_j \frac{p_{ij}}{\sqrt{r_i c_j}} x_j \right\}^2 \\
&= \sum_i r_i \left\{ \sum_j q_{ij} \frac{x_j}{\sqrt{c_j}} \right\}^2 \\
&\leq \sum_i r_i \left\{ \sum_j q_{ij} \left(\frac{x_j}{\sqrt{c_j}} \right)^2 \right\} \\
&= \sum_j x_j^2 \left[\frac{1}{c_j} \sum_i p_{ij} \right] \\
&= \|\mathbf{x}\|^2,
\end{aligned}$$

as required.

- By the above, the corresponding reduced-rank approximation to $\mathbf{P}$ is

$$\begin{aligned}\hat{\mathbf{P}} &= \mathbf{D}_{\mathbf{r}}^{1/2}\hat{\mathbf{B}}\mathbf{D}_{\mathbf{c}}^{1/2} \\ &= \sum_{j=1}^{k} \lambda_j \mathbf{D}_{\mathbf{r}}^{1/2}\mathbf{u}_j\mathbf{v}_j^T\mathbf{D}_{\mathbf{c}}^{1/2} \\ &= \mathbf{rc}^T + \sum_{j=2}^{k} \lambda_j \mathbf{D}_{\mathbf{r}}^{1/2}\mathbf{u}_j\mathbf{v}_j^T\mathbf{D}_{\mathbf{c}}^{1/2}.\end{aligned}$$

  Since $\mathbf{rc}^T$ is common to all such representations, we study $\mathbf{P}_0 = \mathbf{P} - \mathbf{rc}^T$. Note that $\mathbf{P}_0 = \mathbf{0}$ if the two factors are exactly independent. In any event, its row and column sums are all $= 0$.

- Correspondingly,

$$\mathbf{B}_0 = \mathbf{D}_{\mathbf{r}}^{-1/2}\mathbf{P}_0\mathbf{D}_{\mathbf{c}}^{-1/2}$$

  is decomposed (svd) as

$$\mathbf{B}_0 = \mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^T = \sum_{j=1}^{J-1} \tilde{\lambda}_j \tilde{u}_j \tilde{v}_j^T$$

(note $J-1$ – the 'dominant' eigenvalue is gone) and approximated by

$$\hat{\mathbf{B}}_0 = \sum_{j=1}^{k} \tilde{\lambda}_j \tilde{u}_j \tilde{v}_j^T.$$

The inertias are defined, for $j = 1, ..., J-1$, by

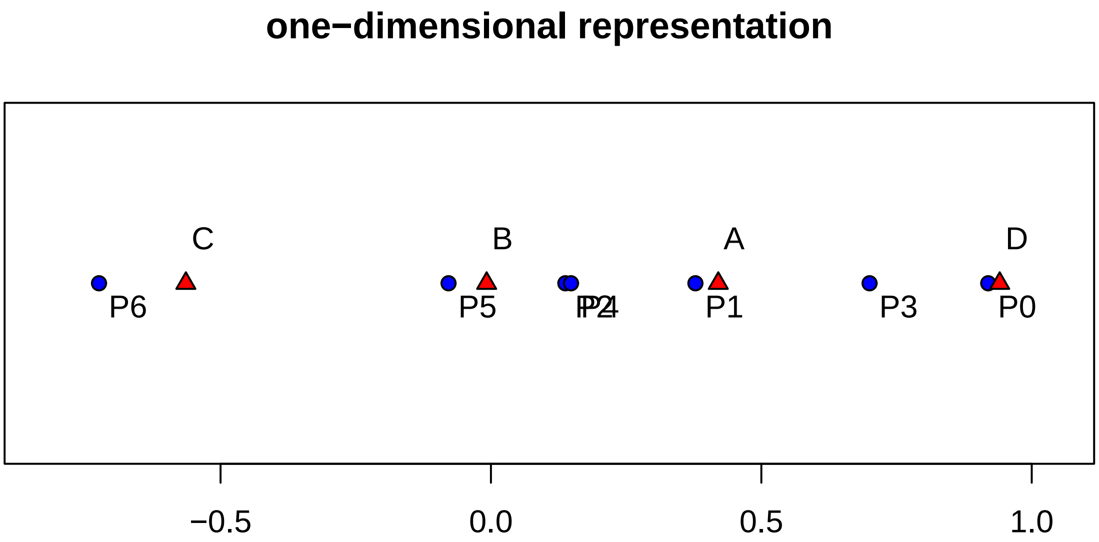
$$inertia_j = \tilde{\lambda}_j^2;$$

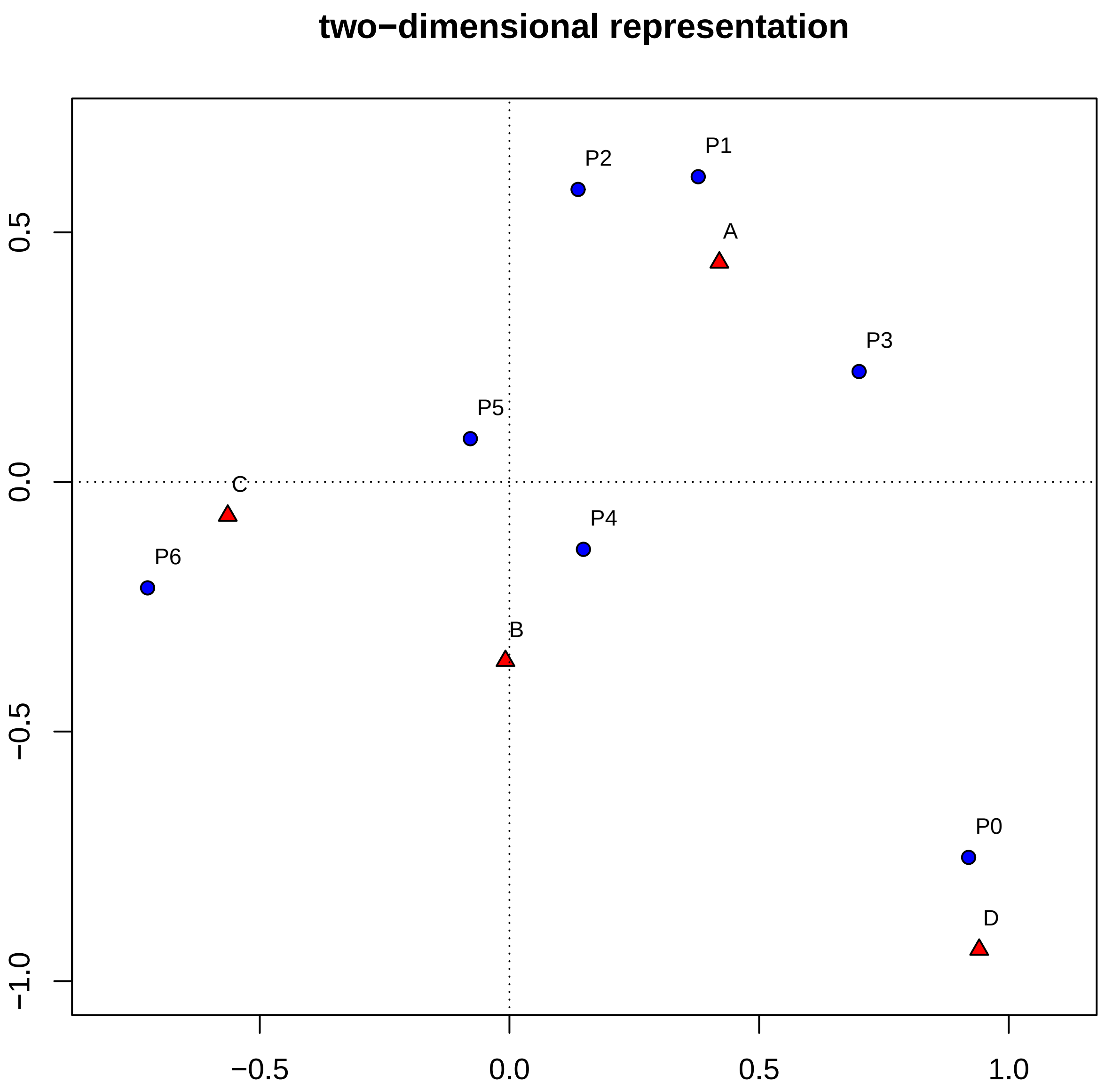
interpretation is as above. Note that

$$\sum_{j=1}^{J-1} inertia_j = \sum_{i,j} [\mathbf{B}_0]_{i,j}^2 = \sum_{i,j} \frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \chi^2/n.$$

- Finally, define $\check{u}_j = \mathbf{D_r}^{-1/2}\tilde{u}_j$ and $\check{v}_j = \mathbf{D_c}^{-1/2}\tilde{v}_j$. A *symmetric map* is a plot of the k-dimensional rows of $\mathbf{F} = \left[\tilde{\lambda}_1 \check{u}_1 \vdots \cdots \vdots \tilde{\lambda}_k \check{u}_k\right]_{I\times k}$ and of $\mathbf{G} = \left[\tilde{\lambda}_1 \check{v}_1 \vdots \cdots \vdots \tilde{\lambda}_k \check{v}_k\right]_{J\times k}$, shown here for $k = 1, 2$. Points close to each other indicate associated attributes. An adequate dimensionality is given by the inertias: for instance if the first two eigenvalues account for a large proportion of the inertia we say that the associations (in the centred data) are well represented by points in a plane.

- Note $\check{u}_\alpha^T \mathbf{D_r} \check{u}_\alpha = \check{v}_\beta^T \mathbf{D_c} \check{v}_\beta = 1$ for $\alpha, \beta = 1, ..., k$; i.e. the weighted averages $\sum_{i=1}^{I} r_i \check{u}_{\alpha,i}^2 = \sum_{j=1}^{J} c_j \check{v}_{\beta,j}^2$ all $= 1$. Thus the weighted sums of squares of the elements in the j^{th} columns of $\mathbf{F}$ and $\mathbf{G}$ are both $= \tilde{\lambda}_j^2$, the contribution to the total inertia. This is the default normalization; others are described in a paper on the web site.




two−dimensional representation
P0
P1
P2
P3
P4
P5
P6
A
B
C
D
−1.0
−0.5
0.0
0.5
−0.5
0.0
0.5
1.0