II. Foundations of Multivariate Analysis

A Some Matrix Algebra

Partitioned Matrices

$$\underset{(m+q)\times(n+r)}{\mathbf{A}} = \begin{bmatrix} \underset{m\times n}{\mathbf{A}_{11}} & \underset{m\times r}{\mathbf{A}_{12}} \\ \underset{q\times n}{\mathbf{A}_{21}} & \underset{q\times r}{\mathbf{A}_{22}} \end{bmatrix} \quad \underset{(n+r)\times(s+t)}{\mathbf{B}} = \begin{bmatrix} \underset{n\times s}{\mathbf{B}_{11}} & \underset{n\times t}{\mathbf{B}_{12}} \\ \\ \underset{r\times s}{\mathbf{B}_{21}} & \underset{r\times t}{\mathbf{B}_{22}} \end{bmatrix}.$$

– Since all of the submatrices are conformable,

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}$$

– If $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square ($m = n$ and $q = r$) and $\mathbf{A}_{12} = \mathbf{A}'_{21} = \mathbf{0}$,

$$|\mathbf{A}| = |\mathbf{A}_{11}|\,|\mathbf{A}_{22}|$$

– If $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are both square and nonsingular,

$$\begin{aligned}
|\mathbf{A}| &= |\mathbf{A}_{11}| \left|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right| \\
&= |\mathbf{A}_{22}| \left|\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right|
\end{aligned}$$

analogous to

$$\begin{aligned}
\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= a_{11}a_{22} - a_{21}a_{12} \\
&= a_{11}\left(a_{22} - \frac{a_{21}a_{12}}{a_{11}}\right) \\
&= a_{22}\left(a_{11} - \frac{a_{12}a_{21}}{a_{22}}\right)
\end{aligned}$$

Eigenvalues and Eigenvectors

λ is an eigenvalue of the square matrix $\mathbf{A}$ and $\mathbf{x}$ is the corresponding eigenvector if:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

- If $|\mathbf{A} - \lambda\mathbf{I}| \neq 0$, then $(\mathbf{A} - \lambda\mathbf{I})$ has an inverse and

$$(\mathbf{A} - \lambda\mathbf{I})^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = (\mathbf{A} - \lambda\mathbf{I})^{-1}\ \mathbf{0}$$

$$\Rightarrow \mathbf{x} = \mathbf{0} \text{ is the only solution}$$

- So, set $\underbrace{|\mathbf{A} - \lambda\mathbf{I}_p| = 0}_{\text{“characteristic equation”}}$ and solve for λ

- Eigenvalues $\lambda_1, \ldots, \lambda_p$ accompanied by eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$.

Orientation of eigenvectors is what's important. Length is arbitrary — $k\mathbf{x}_1, \ldots, k\mathbf{x}_p$ equally good. (We usually choose eigenvectors such that $\mathbf{x}'\mathbf{x} = 1$.)

- Spectral decomposition of $\mathbf{A}$

  Let

$$
\begin{aligned}
\mathbf{C} &= \text{matrix containing normalized eigenvectors of } \mathbf{A} \\
&= [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_p]
\end{aligned}
$$

  and let

$$
\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}
$$

  Note that $\mathbf{C}$ is orthogonal (so $\mathbf{I} = \mathbf{CC}' = \mathbf{C}'\mathbf{C}$)

$$
\begin{aligned}
\mathbf{A} &= \mathbf{ACC}' \\
&= \mathbf{A}[\mathbf{x}_1 \quad \mathbf{x}_2 \quad \ldots \quad \mathbf{x}_p]\mathbf{C}' \\
&= [\mathbf{Ax}_1 \quad \mathbf{Ax}_2 \quad \ldots \quad \mathbf{Ax}_p]\mathbf{C}' \\
&= [\lambda_1\mathbf{x}_1 \quad \lambda_2\mathbf{x}_2 \quad \ldots \quad \lambda_p\mathbf{x}_p]\mathbf{C}' \\
&= \mathbf{CDC}' \\
&= \lambda_1\mathbf{x}_1\mathbf{x}_1' + \lambda_2\mathbf{x}_2\mathbf{x}_2' + \cdots + \lambda_p\mathbf{x}_p\mathbf{x}_p'
\end{aligned}
$$

Also,

$$
\mathbf{D} = \mathbf{C}'\mathbf{AC}
$$

Positive Definite & Nonnegative Definite Matrices

- $\underset{p\times p}{\mathbf{A}}$ is symmetric
- If $\mathbf{x}'\mathbf{A}\mathbf{x} > \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{A}$ is "positive definite (p.d.)"
- If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{A}$ is "nonnegative definite (n.n.d.)"
- If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$, with $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{0}$ for at least one $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{A}$ is "positive semi-definite (p.s.d.)"
  - In other words, if $\mathbf{A}$ is n.n.d., but not p.d., we say $\mathbf{A}$ is p.s.d.

Easy check for these properties:

1. Eigenvalues of a positive definite matrix are all positive.
2. Eigenvalues of a nonnegative definite matrix are positive or zero (with rank($\mathbf{A}$) = number of positive eigenvalues).

Trace and Determinant of a Square Matrix $\underset{p\times p}{\mathbf{A}} = (a_{ij})$

- $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{p} a_{ii}$
- $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$

If $\mathbf{A}$ has e'vals $\lambda_1, \ldots, \lambda_p$

- $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{p} \lambda_i$
- $|\mathbf{A}| = \prod_{i=1}^{p} \lambda_i$ (Practice: show these two statements are true)

Square-root and Inverse matrices

The spectral decomposition of symmetric $\underset{p\times p}{\mathbf{A}}$:

$$\mathbf{A} = \mathbf{CDC}'$$

- $\mathbf{A}^{1/2} = \mathbf{CD}^{1/2}\mathbf{C}'$

$$\text{where } \mathbf{D}^{1/2} = \begin{bmatrix} \lambda_1^{1/2} & & & \mathbf{0} \\ & \lambda_2^{1/2} & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_p^{1/2} \end{bmatrix}$$

$$\text{Note: } \mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{C}\mathbf{D}^{1/2}\mathbf{C}'\mathbf{C}\mathbf{D}^{1/2}\mathbf{C}'$$
$$= \mathbf{C}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{C}'$$
$$= \mathbf{C}\mathbf{D}\mathbf{C}'$$
$$= \mathbf{A}$$

- $\mathbf{A}^{-1} = \mathbf{C}\mathbf{D}^{-1}\mathbf{C}'$

$$\text{where } \mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & & & \mathbf{0} \\ & \frac{1}{\lambda_2} & & \\ & & \ddots & \\ \mathbf{0} & & & \frac{1}{\lambda_p} \end{bmatrix}$$

$$
\begin{aligned}
\text{Note: } \mathbf{A}^{-1}\mathbf{A} &= \mathbf{C}\mathbf{D}^{-1}\mathbf{C}'\mathbf{C}\mathbf{D}\mathbf{C}' \\
&= \mathbf{C}\mathbf{D}^{-1}\mathbf{D}\mathbf{C}' \\
&= \mathbf{C}\mathbf{C}' \\
&= \mathbf{I}
\end{aligned}
$$

- Suppose $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \mathbf{a}'_{12} & a_{22} \end{bmatrix}$

$$\mathbf{A}^{-1} = \frac{1}{b} \begin{bmatrix} b\mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{a}_{12}\mathbf{a}'_{12}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{a}_{12} \\ -\mathbf{a}'_{12}\mathbf{A}_{11}^{-1} & 1 \end{bmatrix}$$

  where $b = a_{22} - \mathbf{a}'_{12}\mathbf{A}_{11}^{-1}\mathbf{a}_{12}$

- Suppose $\underset{p \times p}{\mathbf{A}} = \mathbf{B} + \mathbf{c}\mathbf{c}'$

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} - \frac{\mathbf{B}^{-1}\mathbf{c}\mathbf{c}'\mathbf{B}^{-1}}{1 + \mathbf{c}'\mathbf{B}^{-1}\mathbf{c}}$$

Random Vectors and Matrices

- $\mathbf{X} = (x_{ij})$ is a random matrix (matrix of r.v.'s)

- $E\{\mathbf{X}\} = \begin{bmatrix} E\{x_{11}\} & E\{x_{12}\} & \cdots & E\{x_{1p}\} \\ E\{x_{21}\} & E\{x_{22}\} & \cdots & E\{x_{2p}\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_{n1}\} & E\{x_{n2}\} & \cdots & E\{x_{np}\} \end{bmatrix}$

- If $\mathbf{X}$ and $\mathbf{Y}$ are random and $\mathbf{A}$ and $\mathbf{B}$ are constant:

$$E\{\mathbf{X} + \mathbf{Y}\} = E\{\mathbf{X}\} + E\{\mathbf{Y}\}$$

$$E\{\mathbf{AXB}\} = \mathbf{A}E\{\mathbf{X}\}\mathbf{B}$$

- If $\underset{p\times 1}{\mathbf{x}}$ is a random vector

$$E\{\mathbf{x}\} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \underset{p\times 1}{\boldsymbol{\mu}}$$

$$\boldsymbol{\Sigma} = \mathrm{var}\{\mathbf{x}\} \qquad (\leftarrow \text{ or } \text{“cov}\{\mathbf{x}\}\text{”})$$

$$= E\{(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'\}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \left.\vphantom{\begin{bmatrix} \sigma \\ \sigma \\ \vdots \\ \sigma \end{bmatrix}}\right\}$$

Notes:

1. $\sigma_{ij} = \sigma_{ji}$
2. x_i indep. of x_j $\Rightarrow$ $\mathrm{cov}(x_i, x_j) = 0$

$$\mathbf{P} = \mathrm{corr}\{\mathbf{x}\} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \left.\vphantom{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}\right\}$$

Note:

$\rho_{ij} = \rho_{ji}$

- Let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_q \\ x_{q+1} \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \begin{matrix} \leftarrow \ q \times 1 \\ \leftarrow \ (p-q) \times 1 \end{matrix}$$

$$E\{\mathbf{x}\} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

$$\text{var}\{\mathbf{x}\} = \begin{bmatrix} \underbrace{\boldsymbol{\Sigma}_{11}}_{q\times q} & \underbrace{\boldsymbol{\Sigma}_{12}}_{q\times(p-q)} \\ \underbrace{\boldsymbol{\Sigma}_{21}}_{(p-q)\times q} & \underbrace{\boldsymbol{\Sigma}_{22}}_{(p-q)\times(p-q)} \end{bmatrix}$$

- Let $\underset{r\times p}{\mathbf{B}}$ and $\underset{m\times p}{\mathbf{C}}$ be constant matrices and let $\underset{r\times 1}{\mathbf{b}}$ be $\underset{m\times 1}{\mathbf{c}}$ constant vectors

$$
\begin{aligned}
E\{\mathbf{Bx}\} &= \mathbf{B}\boldsymbol{\mu} \qquad \leftarrow r\times 1 \\
\mathrm{var}\{\mathbf{Bx}\} &= \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}' \qquad \leftarrow r\times r \\
\mathrm{cov}\{\mathbf{Bx},\mathbf{Cx}\} &= \mathbf{B}\boldsymbol{\Sigma}\mathbf{C}' \qquad \leftarrow r\times m \\
E\{\mathbf{b}'\mathbf{x}\} &= \mathbf{b}'\boldsymbol{\mu} \qquad \leftarrow \text{scalar} \\
\mathrm{var}\{\mathbf{b}'\mathbf{x}\} &= \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b} \qquad \leftarrow \text{scalar} \\
\mathrm{cov}\{\mathbf{b}'\mathbf{x},\mathbf{c}'\mathbf{x}\} &= \mathbf{b}'\boldsymbol{\Sigma}\mathbf{c} \qquad \leftarrow \text{scalar}
\end{aligned}
$$

B Expected values for $\bar{\mathbf{x}}$ and $\mathbf{S}$

Let $\mathbf{x}_i$, $i = 1, \ldots, n$, be an i.i.d. random sample with $E\{\mathbf{x}_i\} = \boldsymbol{\mu}$ and $\mathrm{var}\{\mathbf{x}_i\} = \boldsymbol{\Sigma}$.

$$
\begin{aligned}
E\{\bar{\mathbf{x}}\} &= \frac{1}{n}\left(E\{\mathbf{x}_1\} + E\{\mathbf{x}_2\} + \cdots + E\{\mathbf{x}_n\}\right) \\
&= \frac{1}{n}\left(n\boldsymbol{\mu}\right) \\
&= \boldsymbol{\mu}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}\{\bar{\mathbf{x}}\} &= E\{(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\} \\
&=
\end{aligned}
$$

•

Note:

$$\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

$$= \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{x}_i' - \left(\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\right)\bar{\mathbf{x}}'$$

$$= \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' - \sum_{i=1}^{n}\bar{\mathbf{x}}\mathbf{x}_i'$$

$$= \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' - n\bar{\mathbf{x}}\bar{\mathbf{x}}'$$

$$E\{\mathbf{S}\} = E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\right\}$$

$$=$$

$$E\{\mathbf{S}_n\} = E\{\frac{n-1}{n}\mathbf{S}\} = \frac{n-1}{n}\mathbf{\Sigma}$$

C Geometry of the Sample

Vectors

- Length of vector $\mathbf{x} = (x_1, \ldots, x_n)$

$$L_{\mathbf{x}} = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}'\,\mathbf{x}}$$

- For some constant c,

$$L_{(c\mathbf{x})} = |c|\, L_{\mathbf{x}}$$

- Angle between vectors

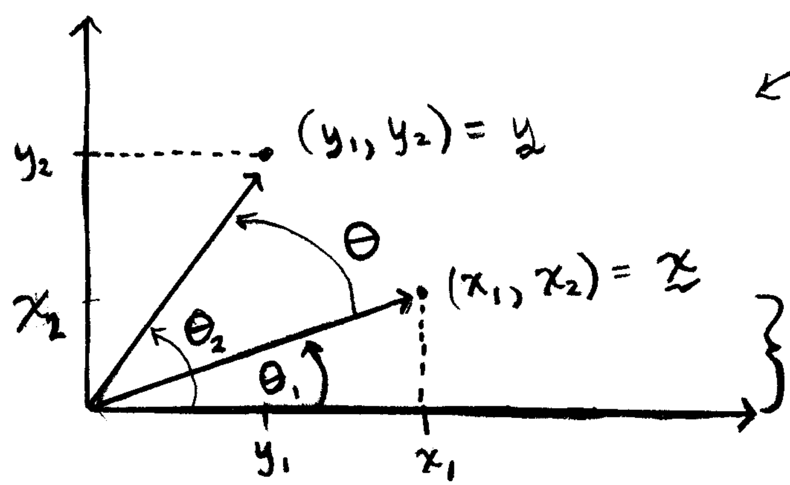


Fig. 2.4 in JW

$x_2 = \sin(\theta_1) \cdot L_x$

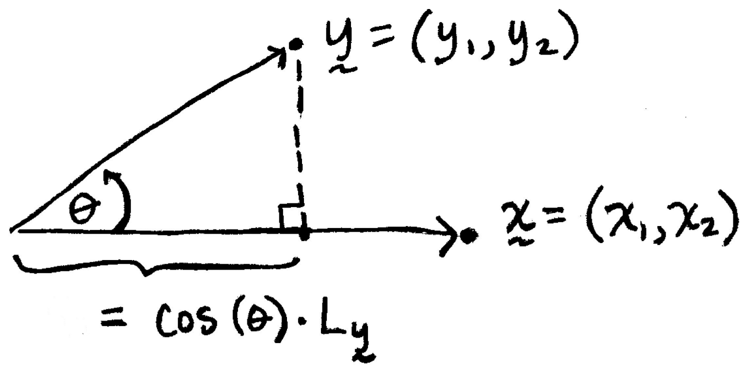
$$
\begin{aligned}
\cos(\theta) &= \cos(\theta_2 - \theta_1) \\
&= \cos(\theta_2)\cos(\theta_1) + \sin(\theta_2)\sin(\theta_1) \\
&= \left(\frac{y_1}{L_y}\right)\left(\frac{x_1}{L_{\mathbf{x}}}\right) + \left(\frac{y_2}{L_y}\right)\left(\frac{x_2}{L_{\mathbf{x}}}\right) \\
&= \frac{\mathbf{x}'\mathbf{y}}{L_{\mathbf{x}}L_{\mathbf{y}}}
\end{aligned}
$$

So

$$\mathbf{x}'\mathbf{y} = 0 \iff \cos(\theta) = 0 \iff \theta \text{ is } 90^\circ \text{ or } 270^\circ$$

Thus $\mathbf{x}'\mathbf{y} = 0$ means $\mathbf{x}$ and $\mathbf{y}$ are perpendicular.

Projections



Projection (shadow) of $\mathbf{y}$ onto $\mathbf{x}$

$$= \frac{\mathbf{y}'\mathbf{x}}{\mathbf{x}'\mathbf{x}}\mathbf{x}$$

Length of projection of $\mathbf{y}$ onto $\mathbf{x}$

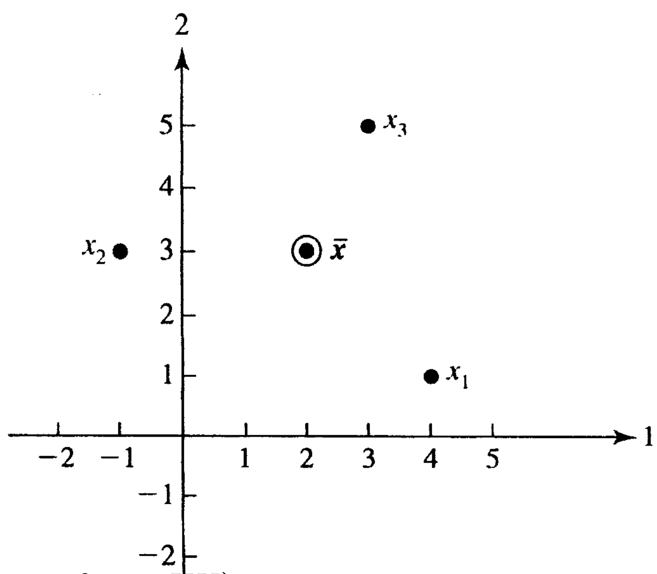
$$= \left|\frac{\mathbf{y}'\mathbf{x}}{\mathbf{x}'\mathbf{x}}\right| L_{\mathbf{x}} = \frac{|\mathbf{y}'\mathbf{x}|}{L_{\mathbf{x}}^2} L_{\mathbf{x}} = \frac{|\mathbf{y}'\mathbf{x}|}{L_{\mathbf{x}}}$$

$$= |\cos(\theta)| \cdot L_{\mathbf{y}} \qquad \text{since } \mathbf{y}'\mathbf{x} = \cos(\theta) L_{\mathbf{y}} L_{\mathbf{x}}$$

Projections of Sample Vectors

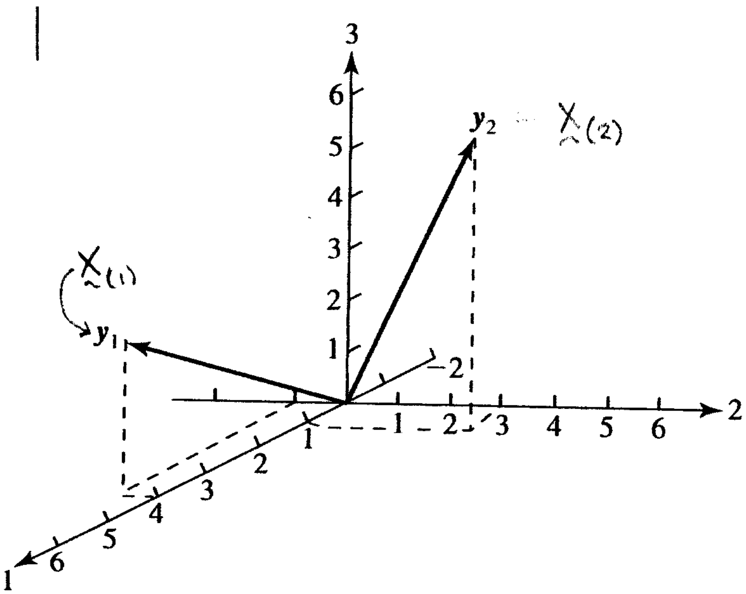
- Viewing $\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$ as n points in p-space,

$$\begin{aligned} \underset{p\times 1}{\bar{\mathbf{x}}} &= \frac{1}{n}\mathbf{1}'\mathbf{X} \\ &= (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)' \\ &= \text{the center of gravity} \end{aligned}$$



(Fig 3.1 from JW)

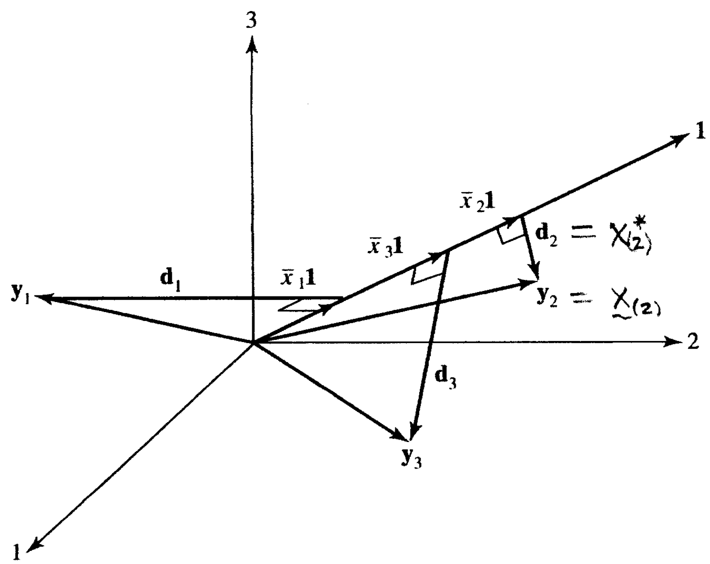
- Alternatively, view $\underset{n\times p}{\mathbf{X}} = \left[\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \cdots, \mathbf{x}_{(p)}\right]$ as p points in n-space



(Fig 3.2 from JW)

  - Projection of $\mathbf{x}_{(i)}$ onto unit-length vector $\frac{1}{\sqrt{n}}\mathbf{1}_n$ is:

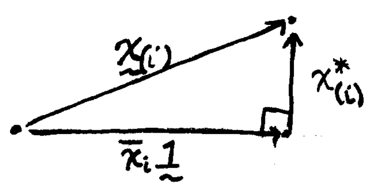
$$\frac{\mathbf{x}'_{(i)}\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)}{\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)'\left(\frac{1}{\sqrt{n}}\mathbf{1}\right)}\frac{1}{\sqrt{n}}\mathbf{1} = \left(\frac{1}{n}\sum_{j=1}^{n} x_{ji}\right)\mathbf{1} = \bar{x}_i\mathbf{1}$$



(Fig 3.3 from JW)

– Note: Centered (mean-corrected) version of $\mathbf{x}_{(i)}$ (also called "deviation vector") is

$$\mathbf{x}^*_{(i)} = \mathbf{x}_{(i)} - \bar{x}_i\mathbf{1}$$



– Note:

$$
\begin{aligned}
\text{Length of } \mathbf{x}^*_{(i)} &= \sqrt{\mathbf{x}^{*'}_{(i)}\mathbf{x}^*_{(i)}} \\
&= \sqrt{\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)^2} \\
&= \sqrt{(n-1)s_{ii}} \\
&= \sqrt{n-1}s_i \\
\text{OR} \qquad s_{ii} &= \frac{1}{n-1}L^2_{\mathbf{x}^*_{(i)}} \\
\text{Similarly,} \qquad s_{ij} &= \frac{1}{n-1}\mathbf{x}^{*'}_{(i)}\mathbf{x}^*_{(j)}
\end{aligned}
$$

- Since, in general, $\mathbf{y}'\mathbf{z} = L_{\mathbf{y}} L_{\mathbf{z}} \cos(\theta)$, it follows that

$$
\begin{aligned}
\cos(\boldsymbol{\theta}) &= \frac{\mathbf{x}_{(i)}^{*'} \mathbf{x}_{(j)}^{*}}{L_{\mathbf{x}_{(i)}^{*}} L_{\mathbf{x}_{(j)}^{*}}} \\
&= \frac{(n-1)s_{ij}}{\sqrt{(n-1)s_{ii}}\sqrt{(n-1)s_{jj}}} \\
&= \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \\
&= r_{ij}
\end{aligned}
$$

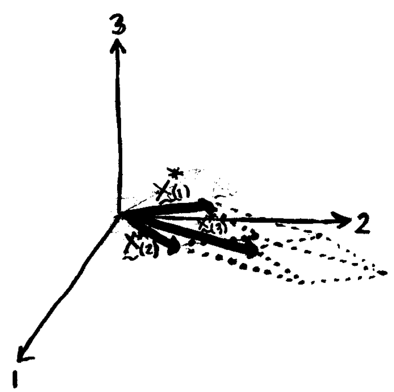
$\therefore$ correlation i^{th} and j^{th} variables is cosine of angle between $\mathbf{x}_{(i)}^{*}$ and $\mathbf{x}_{(j)}^{*}$

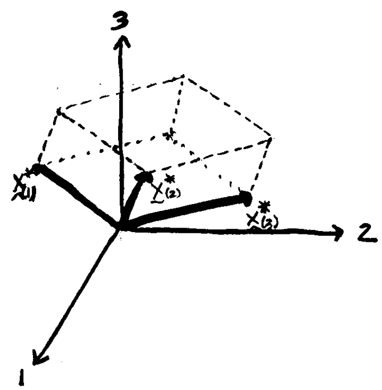
D Generalized Variance

Desire a single value which summarizes variability of multivariate observations.

- Recall: $\mathbf{S}$ is a function of deviation vectors $\mathbf{x}^*_{(1)}, \cdots, \mathbf{x}^*_{(p)}$

- It can be shown that $|\mathbf{S}| = \frac{(\text{volume})^2}{(n-1)^p}$
  where volume is the p-dimensional volume of the p-dimensional "box" formed by $\mathbf{x}^*_{(1)}, \cdots, \mathbf{x}^*_{(p)}$





- $|\mathbf{S}|$ is "<u>generalized sample variance</u>"

- $|\mathbf{S}|$ larger as $\mathbf{x}^*_{(1)}, \cdots, \mathbf{x}^*_{(p)}$ are re-oriented to be nearly perpendicular (without changing lengths)

- $|\mathbf{S}|$ larger when $\mathbf{x}^*_{(i)}$ is increased in length ($\mathbf{x}_{(i)}$ multiplied by $c > 1$) without changing orientation

- $|\mathbf{S}| \cong 0$ when any $\mathbf{x}^*_{(i)} \cong \mathbf{0}$ (i.e., small s_{ii})

- $|\mathbf{S}| \cong 0$ when any $\mathbf{x}^*_{(i)}$ lies nearly in $(p-1)$-dim. hyper-plane formed by other deviation vectors

$$\mathbf{x}^*_{(i)} \cong a_1\mathbf{x}^*_{(1)} + \cdots + a_{i-1}\mathbf{x}^*_{(i-1)} + a_{i+1}\mathbf{x}^*_{(i+1)} + \cdots + a_p\mathbf{x}^*_{(p)}$$

- $|\mathbf{S}| = 0$ if one or more of observed variables is a linear function (sum, difference, etc.) of one or more other observed variables

- $|\mathbf{R}| = \frac{(\text{volume})^2}{(n-1)^p}$ where volume is formed by standardized deviation vectors $\frac{\mathbf{x}^*_{(1)}}{\sqrt{s_{11}}}, \cdots, \frac{\mathbf{x}^*_p}{\sqrt{s_{pp}}}$

- $|\mathbf{R}| = \frac{1}{(s_{11}, s_{22}, \cdots, s_{pp})}|\mathbf{S}|$

- $|\mathbf{R}|$ unaffected by multiplying $\mathbf{x}_{(i)}$ by $c \neq 0$
- If $s_{(ij)} = 0$ for all $i \neq j$, $|\mathbf{R}| = 1$
- Alternative to $|\mathbf{S}|$ = "generalized sample variance" is $\text{tr}\{\mathbf{S}\}$ = "total sample variance"

  Recall that $|\mathbf{S}| = \prod_{i=1}^{p} \lambda_i$

  $$\text{tr}\{\mathbf{S}\} = \sum_{\mathbf{i=1}}^{\mathbf{p}} \lambda_{\mathbf{i}}$$

  $\text{tr}\{\mathbf{S}\}$ incorporates no multivariate (correlation structure) information

E Multivariate Normal (MVN) Distribution

Why the emphasis on the MVN?

(1) Only 1^{st} and 2^{nd} moments needed to describe distribution

(2) Uncorrelated variables $\Rightarrow$ independent variables

(3) Linear functions of MVN variables are normal

(4) Genuinely good population model for some natural phenomena

(5) *Even for nonnormal data, MVN is often useful approximation – especially for inferences involving sample mean vectors, which are asymptotically normal due to CLT*

- The Gaussian (normal) density function
  - Univariate Gaussian (normal) density:

$$
\begin{aligned}
f_x(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \\
&= \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)}
\end{aligned}
$$

– Bivariate case

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{22} \end{pmatrix} \right]$$

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \\ &\quad \times \exp\left\{ \frac{-1}{2(1-\rho_{12}^2)} \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right)^2 - \right.\right. \\ &\quad \left.\left. 2\rho_{12} \left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right) \right] \right\} \end{aligned}$$

∗ For bivariate case, if $\rho_{12} = 0$, $f_{\mathbf{x}}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) = f_{x_1}(x_1) \cdot f_{x_2}(x_2)$

– p-variate normal density

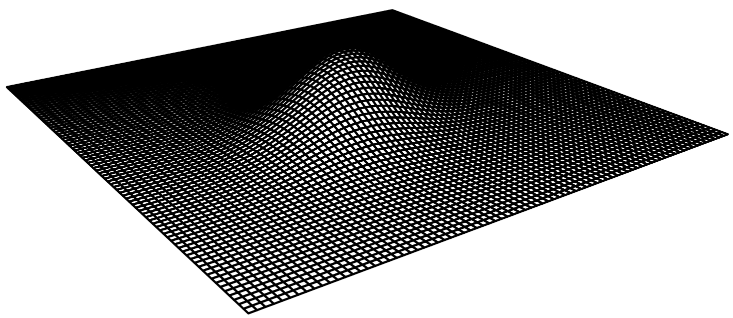
$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

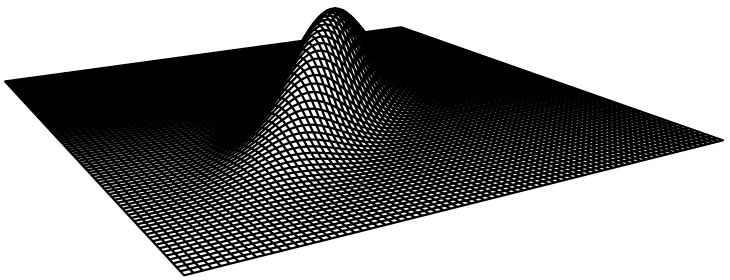
* For p-variate case, if $\mathbf{\Sigma}$ is diagonal

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_{11}} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{\sigma_{pp}} \end{bmatrix} \text{ and } |\mathbf{\Sigma}| = (\sigma_{11})(\sigma_{22})\cdots(\sigma_{pp})$$

$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{(\sigma_{11})\cdots(\sigma_{pp})}} \\ &\quad \times \exp\left\{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_{11}} - \cdots - \frac{1}{2}\frac{(x_p-\mu_p)^2}{\sigma_{pp}}\right\} \\ &= f_{x_1}(x_1)\cdot f_{x_2}(x_2)\cdot \ \cdots \ \cdot f_{x_p}(x_p) \end{aligned}$$

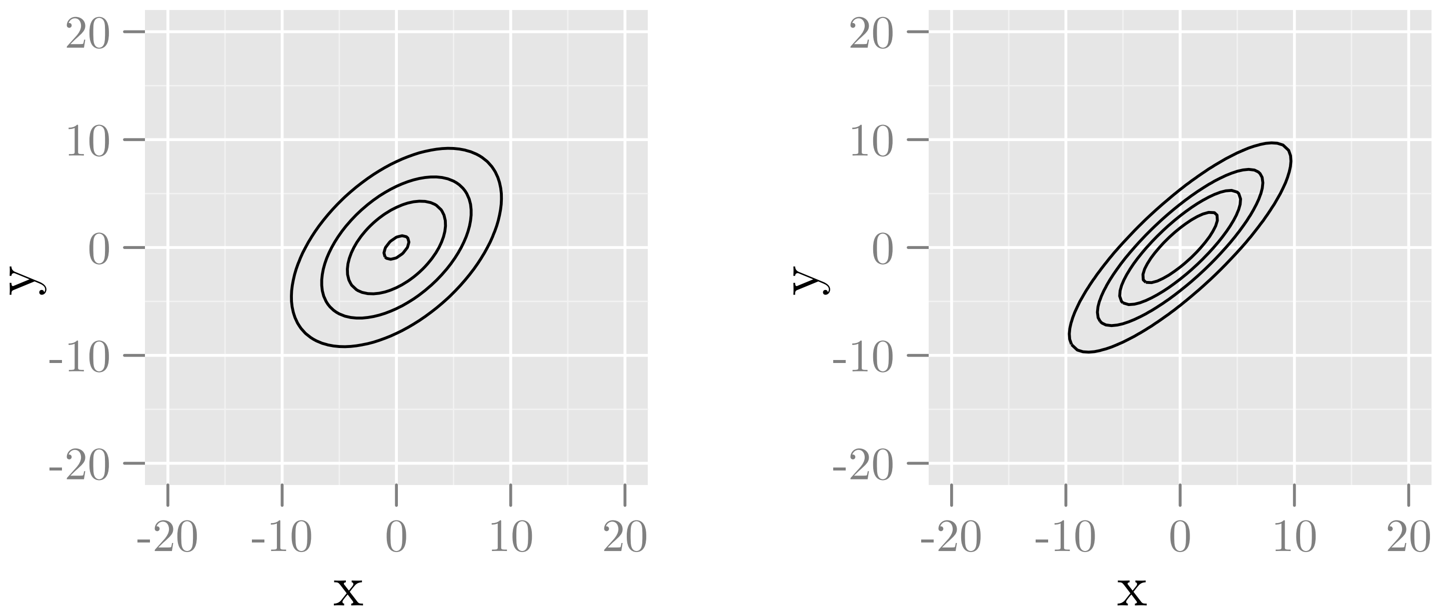
- Shape of the MVN density





(Fig 4.2 from RC)

$\sigma_{11} = \sigma_{22}, \rho_{12} = 0$ $\sigma_{11} = \sigma_{22}, \rho_{12} = .75$



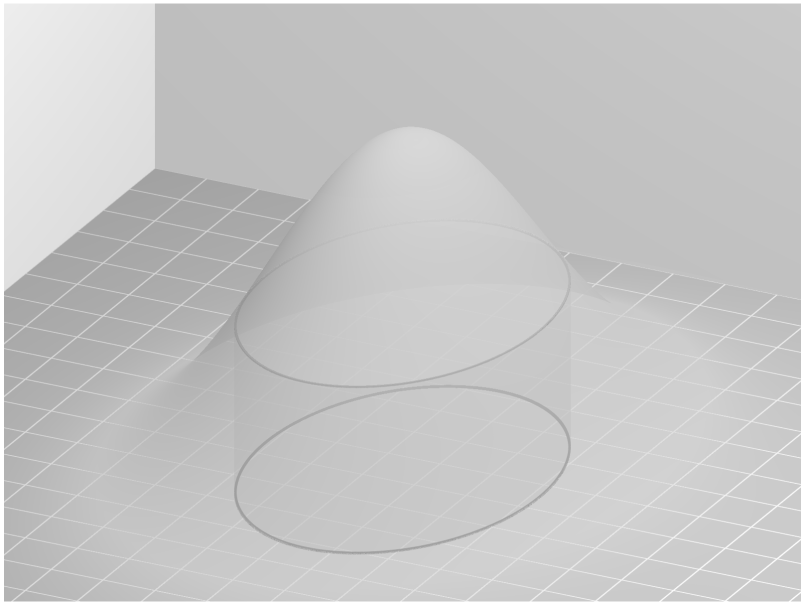
(Fig 4.3 from RC)

$\sigma_{11} = \sigma_{22}$ for both plots

– Which has small $|\mathbf{\Sigma}|$ and which has large $|\mathbf{\Sigma}|$?

- Contours of MVN Values of $\mathbf{x}$ yielding constant height for density are ellipsoids.

  Constant probability density contour
  $= \left\{\text{all } \mathbf{x} \ni (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu) = c^2\right\}$



(Constant density contour for bivariate normal. Fig 4.4 from RC)

  – $\Pr\left\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \leq \chi^2_p(\alpha)\right\} = 1-\alpha$
  where $\chi^2_p(\alpha)$ is the upper (100α)th %-ile

Some Properties of the MVN Distribution

$$\underset{p\times 1}{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ and } \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_p\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

1. Linear combinations of $\mathbf{x}$ are normal

   For constant vector $\underset{q\times 1}{\mathbf{c}}$ and matrix $\underset{q\times p}{\mathbf{A}}$

   - $\mathbf{A}\mathbf{x} + \mathbf{c} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
   - $\mathbf{c}'\mathbf{x} \sim N_1(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$
   - $E\{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})\} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu} - \boldsymbol{\mu}) = \mathbf{0_p}$
     $\text{var}\{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})\} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I_p}$
     and $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I})$
   - $(\mathbf{T}')^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I})$ where $\mathbf{T}'\mathbf{T}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$.

2. All subsets of components of $\mathbf{x}$ are MVN

   If $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim N_p \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$,

$$\begin{aligned} \text{then} \quad \mathbf{x}_1 &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ \mathbf{x}_2 &\sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \\ x_i &\sim N_1(\mu_i, \sigma_{ii}), \quad i = 1, \ldots, p \end{aligned}$$

   QUESTION: Is the converse also true? I.e., if each $x_i,\ i = 1, ..., p$, is distributed normally, does that imply that $\underset{p\times 1}{\mathbf{x}}$ is MVN?

3. Zero covariance $\Leftrightarrow$ independence

   - $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$
   - x_i and x_j are independent if $\sigma_{ij} = 0$

4. Conditional distributions are normal

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

- $E\{\mathbf{x}_1|\mathbf{x}_2\}$ indicates linear relationship between subsets of $\mathbf{x}$ or between x_i and x_j
  - Use to check for nonnormality in bivariate (or p-variate) data

5. Chi-square distribution

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_p$$

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \underbrace{\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})\right]'}_{\underbrace{\mathbf{z}'}_{\sim N_p(\mathbf{0},\mathbf{I})}} \underbrace{\left[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x}-\boldsymbol{\mu})\right]}_{\underbrace{\mathbf{z}}_{\sim N_p(\mathbf{0},\mathbf{I})}}$$

$$= \sum_{i=1}^{p} z_i^2 \quad \text{(sum of } p \text{ indep. squared normals)}$$

$$= \chi^2_p$$

(we'll use this property to check for MVN'ity)

F Assessing MVN'ity & Detecting Outliers

- Though normality of univariate & bivariate subsets of $\underset{p\times 1}{\mathbf{x}}$ does not guarantee MVN'ity, in practice, 1-D and 2-D investigations are often sufficient

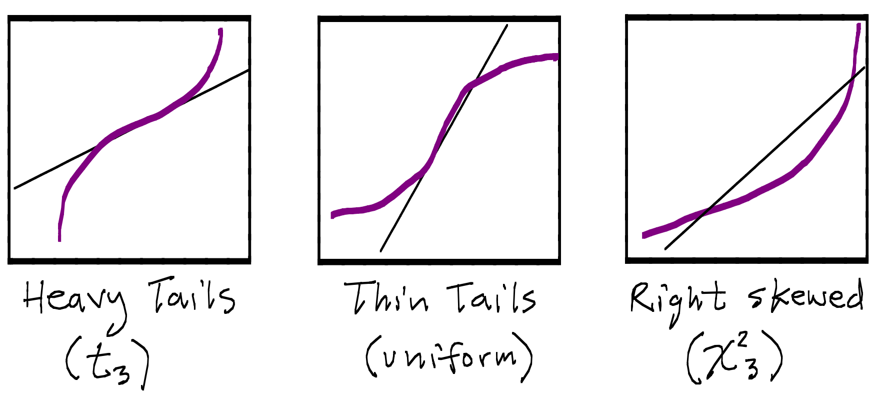
  <u>1-D Tools</u>

  – Histograms

  – Normal probability plots

    y-axis: ordered observations $x_{(1)}, \ldots, x_{(n)}$

    x-axis: $\Phi^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$ or $\Phi^{-1}\left(\frac{i}{n+1}\right)$



– Tests for skewness & kurtosis

– Kolmogorov-Smirnov, D'Agostino, and friends

2-D Tools

– 2-D Scatterplots (check for linearity)

– Check bivariate densities

  * Image plots

  * Perspective plots

Multivariate Tools

– χ^2 QQ-Plot

Since $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ implies $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_p$

Plot:

x-axis: $\left(\frac{i-\frac{1}{2}}{n}\right)^{th}$ quantile of χ^2_p

y-axis: $D^2_{(i)} = i^{th}$ ordered value of D^2_i where

$$D^2_i = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$$

* Alternatively, Gnanadesikan and Kettenring (1972) suggest that the following plot is superior:

  x-axis: $\left(\frac{i-\frac{1}{2}}{n}\right)^{th}$ quantile of $\beta(\frac{p}{2}, \frac{1}{2}(n-p-1))$

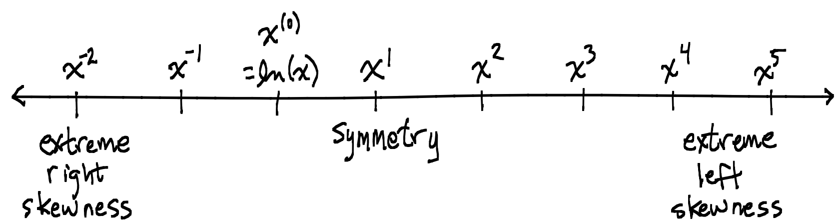
  y-axis: $\frac{n}{(n-1)^2} D^2_{(i)}$

– "Grand Tour"

Univariate Transformations to Near-Normality

- Make data more "normal" by considering various transformations
- Some standard transformations
  * Counts (x) $\Rightarrow$ use $\sqrt{x}$
  * Proportions ($\hat{p}$) $\Rightarrow$ use $\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$
  * Correlations (r) $\Rightarrow$ use $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$
  * Skewed (continuous) data (x) $\Rightarrow$ use "power transformation" (x^λ) or "Box-Cox transformation"

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(x) & \text{for } \lambda = 0 \end{cases}$$



- Box and Cox (1964) recommend using

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(x) & \text{for } \lambda = 0 \end{cases}$$

where λ is chosen by maximizing

$$\ell(\lambda) = -\frac{n}{2} \ln s_\lambda^2 + (\lambda - 1) \sum_{i=1}^{n} \ln(x_i),$$

where

$$s_\lambda^2 = 1/n \sum_{i=1}^{n} (x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2$$

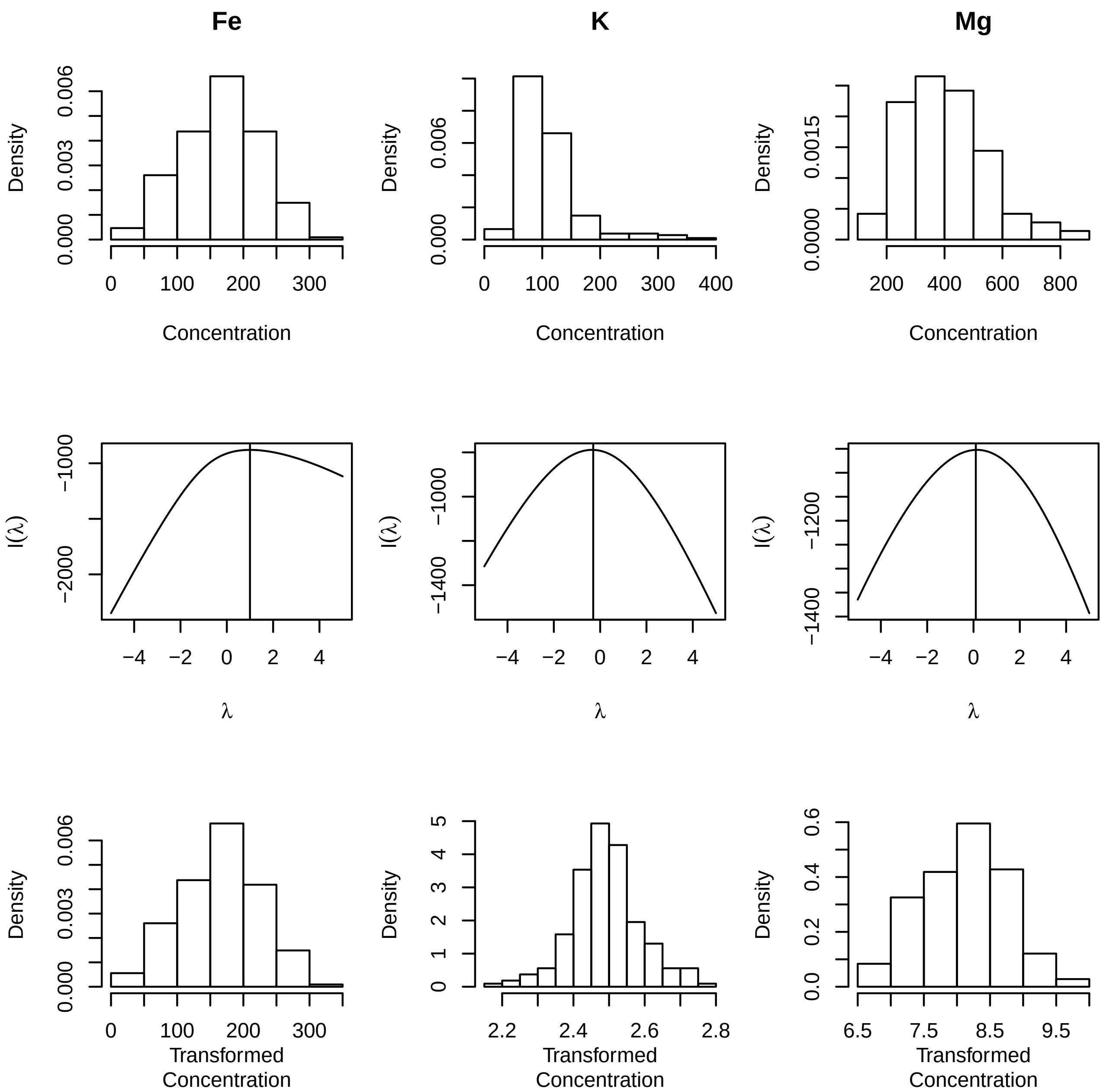
is the maximum likelihood estimate of the variance of $x^{(\lambda)}$ and $\overline{x^{(\lambda)}}$ is the sample mean of the n transformed observations

Multivariate Transformations to Near-Normality

- Maximize

$$\ell(\boldsymbol{\lambda}) = -\frac{n}{2}\ln|\mathbf{S}_{\boldsymbol{\lambda}}| + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}\ln(x_{ij})\right]$$

where x_{ij} is the ith measurement on the jth variable, $\mathbf{S}_{\boldsymbol{\lambda}}$ is the maximum likelihood estimate of the covariance matrix for the transformed data



Fe
K
Mg
Density
0.000
0.003
0.006
0
100
200
300
Concentration
0.006
400
0.0000
0.0015
600
800
l(λ)
−1000
−2000
−4
−2
2
4
λ
−1400
−1200
Transformed
5
0.2
0.4
0.6
2.2
2.4
2.6
2.8
6.5
7.5
8.5
9.5

G Maximum Likelihood

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a r.s. from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Joint density:

$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^n f(\mathbf{x}_i)$

$$= \prod_{i=1}^n \frac{1}{(2\pi)^{\frac{p}{2}}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

$$= \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_i^n (\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}$$

$$= L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Goal: Find values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize the likelihood of observing $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Some preliminaries

- Result 4.10 (Proof on pages 170-171, JW) Given a $p \times p$ symmetric positive definite (p.d.) matrix $\mathbf{B}$ and a scalar $b > 0$,

$$\frac{1}{|\boldsymbol{\Sigma}|^b} e^{\frac{-\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B})}{2}} \leq \frac{1}{|\mathbf{B}|^b}(2b)^{pb} e^{-bp}$$

  for all p.d. $\boldsymbol{\Sigma}$, with equality holding only if $\boldsymbol{\Sigma} = \frac{1}{2b}\mathbf{B}$.

- Rewrite exponent of $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\begin{aligned}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) &= \mathrm{tr}\left\{(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\} \\ &= \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\right\}\end{aligned}$$

  and

$$\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) = \operatorname{tr}\left\{\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right\}$$

$$= \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})'\right\}$$

$$= \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\left[(\mathbf{x}_i-\bar{\mathbf{x}})+(\bar{\mathbf{x}}-\boldsymbol{\mu})\right]\left[(\mathbf{x}_i-\bar{\mathbf{x}})+(\bar{\mathbf{x}}-\boldsymbol{\mu})\right]'\right\}$$

$$= \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'+\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\bar{\mathbf{x}}-\boldsymbol{\mu})' + \sum_{i=1}^{n}(\bar{\mathbf{x}}-\boldsymbol{\mu})(\mathbf{x}_i-\bar{\mathbf{x}})'+\sum_{i=1}^{n}(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right]\right\}$$

$$= \operatorname{tr}\left\{\boldsymbol{\Sigma}^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'\right]\right\}$$

So,

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} \times$$
$$\exp\left\{ -\frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \right] \right) \right\}$$

Note that the value of $\boldsymbol{\mu}$ maximizing $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the value minimizing $\mathrm{tr}\{n\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\}$.

$$\begin{aligned} \mathrm{tr}\left\{ n\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \right\} &= n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \\ &\geq 0 \end{aligned}$$

since $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}$ are p.d. with equality (minimization) when $\bar{\mathbf{x}} = \boldsymbol{\mu}$.

$\therefore$ MLE for $\boldsymbol{\mu}$ is $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$

$$
\begin{aligned}
L(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\ \shortparallel\ )'\right]\right\} \\
&= k\frac{1}{|\boldsymbol{\Sigma}|^b} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{B}\right]\right\} \\
&\qquad\qquad (\text{using Result 4.10, where } b = \tfrac{n}{2} \\
&\qquad\qquad \text{and } \mathbf{B} = \textstyle\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})') \\
&\leq k\frac{1}{|\mathbf{B}|^b}(2b)^{pb}\exp\{-bp\}
\end{aligned}
$$

with equality (maximization) when $\boldsymbol{\Sigma} = \frac{1}{2b}\mathbf{B}$.

$$
\begin{aligned}
\therefore \text{MLE for } \boldsymbol{\Sigma} \text{ is } \widehat{\boldsymbol{\Sigma}} &= \frac{1}{2b}\mathbf{B} \\
&= \frac{1}{n}\Sigma(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \\
&= \mathbf{S}_n
\end{aligned}
$$

Notes:

1. Invariance property: MLE of $h(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $h(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$
2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a r.s. from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\bar{\mathbf{x}}$ and $\mathbf{S}$ are sufficient statistics.

H Sampling Distribution of $\bar{\mathbf{x}}$ and $\mathbf{S}$

Recall for $p = 1$:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

or

$$(n-1)s^2 \sim \sigma^2 \chi^2_{n-1}$$

For $p > 1$:

$$(n-1)\mathbf{S} \sim W_p(n-1, \boldsymbol{\Sigma})$$

$$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

and $\bar{\mathbf{x}}$ and $\mathbf{S}$ are independent

Law of large numbers

$\bar{\mathbf{x}}$ converges in probability to $\boldsymbol{\mu}$

$\mathbf{S}$ converges is probability to $\boldsymbol{\Sigma}$

CLT: Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent obs. from a population with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

- $\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})$ is approx. $N_p(\mathbf{0}, \boldsymbol{\Sigma})$
- $n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$ is approx χ^2_p for $n - p$ large.

I EM Algorithm & Missing Data

Frequently observed scenario:
Many observations contain information on only some of the variables.

Approaches:

1. Analyze only the complete observations
   - May lose substantial amount of data
     – Suppose a mechanism causes $m\%$ of elements of $\underset{n\times p}{\mathbf{X}}$ to be missing at random.

P	10	20	50	100
% of rows of **x** that are complete when m%=1%	90%	82%	61%	37 %
complete rows when m%=5%	60%	36%	8%	0.6%

2. Conduct analysis using

$$\ddot{\bar{\mathbf{x}}} = (\ddot{\bar{x}}_1, \ldots, \ddot{\bar{x}}_p)'$$

and

$$\ddot{\mathbf{S}} = \begin{bmatrix} \ddot{s}_{11} & \cdots & \ddot{s}_{1p} \\ \vdots & \ddots & \vdots \\ \ddot{s}_{p1} & \cdots & \ddot{s}_{pp} \end{bmatrix}$$

- $\ddot{\bar{x}}_i$ calculated using all subjects for which variable i is observable
- $\ddot{s}_{ij}$ calculated using all subjects for which variables i and j are observable
- $\ddot{\mathbf{S}}$ may not be nonnegative def!

3. Replace missing value x_{ij} with $\bar{x}_j$

- Resulting $\mathbf{S}$ is positive definite but each element suffers from attenuation ("shrunk-towards-zero") bias

4. EM Algorithm

- Assumes "missing at random"
  - Mechanism responsible for missingness *not* influenced by value of the variables

    $\boxed{Ex}$ Movie preference data

$$\underset{2000\times 1}{\mathbf{x}_i} = [\underset{\substack{\uparrow \\ \text{"Cinema} \\ \text{Paradiso"}}}{\text{rating}_1}, \ldots, \underset{\substack{\uparrow \\ \text{"The Pokémon} \\ \text{Movie"}}}{\text{rating}_{2000}}]$$

$$\underset{n\times 2000}{\mathbf{x}} = \begin{bmatrix} 7 & 3 & \cdots & \text{NA} \\ \text{NA} & 5 & \cdots & 10 \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

Missing at random??

Two steps of Algorithm:

1. Expectation (Prediction) Step

   Given an estimate $\tilde{\boldsymbol{\theta}}$ $\left(\text{e.g., } \tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\mu}}', \text{vech } \tilde{\boldsymbol{\Sigma}})\right)$, predict the contribution of any missing observation to the (complete-data) sufficient statistics using complete data & current $\tilde{\boldsymbol{\theta}}$.

   $$\text{Let } \underset{p\times 1}{\mathbf{x}_i} = \begin{bmatrix} \mathbf{x}_i^{(1)} \\ \mathbf{x}_i^{(2)} \end{bmatrix} \begin{array}{l} \leftarrow \text{missing components } \ (q \times 1) \\ \leftarrow \text{observed components } \ ((p-q) \times 1) \end{array}$$

   $$\underset{p\times 1}{\tilde{\boldsymbol{\mu}}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}^{(1)} \\ \tilde{\boldsymbol{\mu}}^{(2)} \end{bmatrix} \text{ and } \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{bmatrix}$$

   Each "E" step estimates $\mathbf{x}_i^{(1)}$ using regression:

   $$\underset{q\times 1}{\tilde{\mathbf{x}}_i^{(1)}} = \underset{q\times 1}{\tilde{\boldsymbol{\mu}}_i^{(1)}} + \underset{q\times(p-q)}{\mathbf{B}} \left(\underbrace{\mathbf{x}_i^{(2)} - \tilde{\boldsymbol{\mu}}_i^{(2)}}_{(p-q)\times 1} \right)$$

where regression coefficients

$$\mathbf{B} = \underset{q\times(p-q)}{\tilde{\mathbf{\Sigma}}_{12}} \underset{(p-q)\times(p-q)}{\tilde{\mathbf{\Sigma}}_{22}^{-1}}$$

2. Maximization (Estimation) step

   After obtaining new sufficient statistics (from prediction of missing values in E step), obtain revised version of $\tilde{\boldsymbol{\theta}}$.

   - Iterate "E" and "M" steps until convergence
   - Each iteration has guaranteed increase in likelihood ... at very least, we get a local maximum.

J Multiple Imputation

References: Rubin (1987), van Ginkel and Kroonenberg (2014)

Imputing with EM Algorithm

- Too optimistic — assumes that missing observations are perfectly predictable using observed variables
- Need to account for the uncertainty in predicting missing components ($\mathbf{x}_i^{(1)}$) from observed components ($\mathbf{x}_i^{(2)}$)

Multiple Imputation

- Suppose that after convergence with the EM algorithm, we have estimates $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$
- For $m = 1, \ldots, M$ different imputations, obtain a random prediction of the complete data and denote it $\mathbf{X}_{[m]}$
  - $\mathbf{X}_{[m]}$ is obtained by predicting missing values in the ith row as:

$$\underset{q\times 1}{\mathbf{x}_{i,[m]}^{(1)}} = \underset{q\times 1}{\boldsymbol{\mu}_i^{*(1)}} + \underset{q\times(p-q)}{\mathbf{B}^*}\left(\underbrace{\mathbf{x}_i^{(2)} - \boldsymbol{\mu}_i^{*(2)}}_{(p-q)\times 1}\right) + \mathbf{e}_{i,[m]}^{(1)}$$

where $\mathbf{e}_{i,[m]}^{(1)}$ is a draw from a $\mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_{11}^* - \boldsymbol{\Sigma}_{12}^*\boldsymbol{\Sigma}_{22}^{*-1}\boldsymbol{\Sigma}_{21}^*)$

* In R, if `miss` is a Boolean vector indicating the missing locations in the row and `Sig` is $\mathbf{\Sigma}^*$, the covariance matrix for the draw of the N_q vector is:

```
Sig[miss,miss] - Sig[miss,!miss] %*%
solve(Sig[!miss,!miss]) %*% Sig[!miss,miss]
```

- Collect M random estimates of the complete data and denote these $\mathbf{X}_{[m]}, m = 1, \ldots, M$, and from each matrix, obtain the parameter estimate of interest $\underset{k\times 1}{\hat{\boldsymbol{\beta}}_m}$.
  - E.g., $\hat{\boldsymbol{\beta}}_m$ may be just the sample mean $\bar{\mathbf{x}}_m$ (and $k = p$)
  - Note that you want $M >> k$

- Our MI-based estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m$$

- For inference, we use

$$\underset{k\times k}{\widehat{\mathrm{var}}}(\hat{\boldsymbol{\beta}}_{MI}) = \bar{\mathbf{V}} + (1 + \frac{1}{M})\mathbf{B}$$

where

$$\underset{k\times k}{\bar{\mathbf{V}}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_m)$$

and

$$\underset{k\times k}{\mathbf{B}} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_{MI})(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_{MI})'$$

– Note: if $\boldsymbol{\beta} = \boldsymbol{\mu}$ and $\hat{\boldsymbol{\beta}}_m = \bar{\mathbf{x}}_m$, then

$$\hat{\boldsymbol{\mu}}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \bar{\mathbf{x}}_m$$

$$\underset{p \times p}{\widehat{\text{var}}(\hat{\boldsymbol{\mu}}_{MI})} = \bar{\mathbf{V}} + (1 + \frac{1}{M})\mathbf{B}$$

where

$$\underset{p \times p}{\bar{\mathbf{V}}} = \frac{1}{Mn} \sum_{m=1}^{M} \mathbf{S}_m$$

and

$$\underset{p \times p}{\mathbf{B}} = \frac{1}{M-1} \sum_{m=1}^{M} (\bar{\mathbf{x}}_m - \hat{\boldsymbol{\mu}}_{MI})(\bar{\mathbf{x}}_m - \hat{\boldsymbol{\mu}}_{MI})'$$

– Note: some have recommended as an improved estimate of $\text{var}(\hat{\boldsymbol{\beta}}_{MI})$ to use:

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_{MI}) = \bar{\mathbf{V}} + (1 + \frac{1}{M})[tr(\mathbf{B}\bar{\mathbf{V}}^{-1})/k]\bar{\mathbf{V}}$$

– Warning: Hypothesis tests of $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ or $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ will be anti-conservative if standard df formulae are used (especially when the rate of missingness in the data is high). See Rubin (1987) or van Ginkel and Kroonenberg (2014) for additional details on df adjustments.