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:

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

– 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

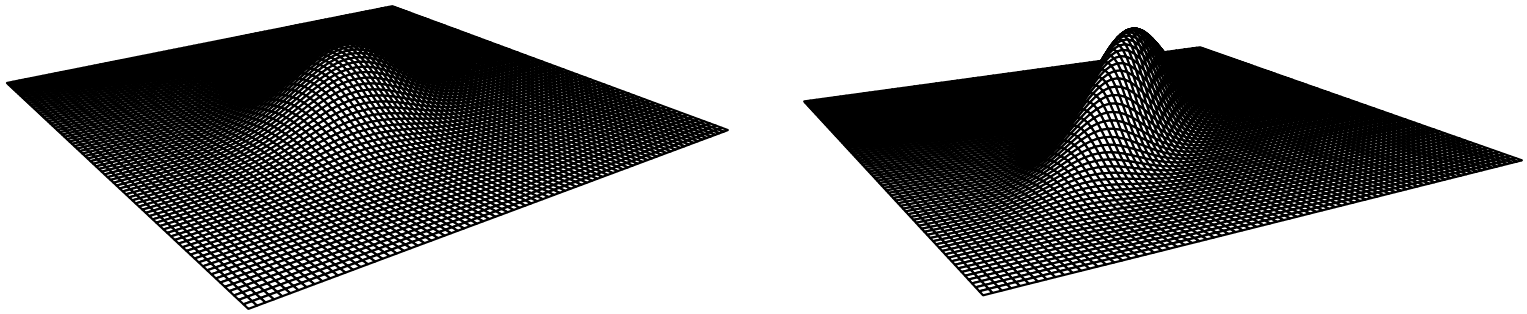
$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

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

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

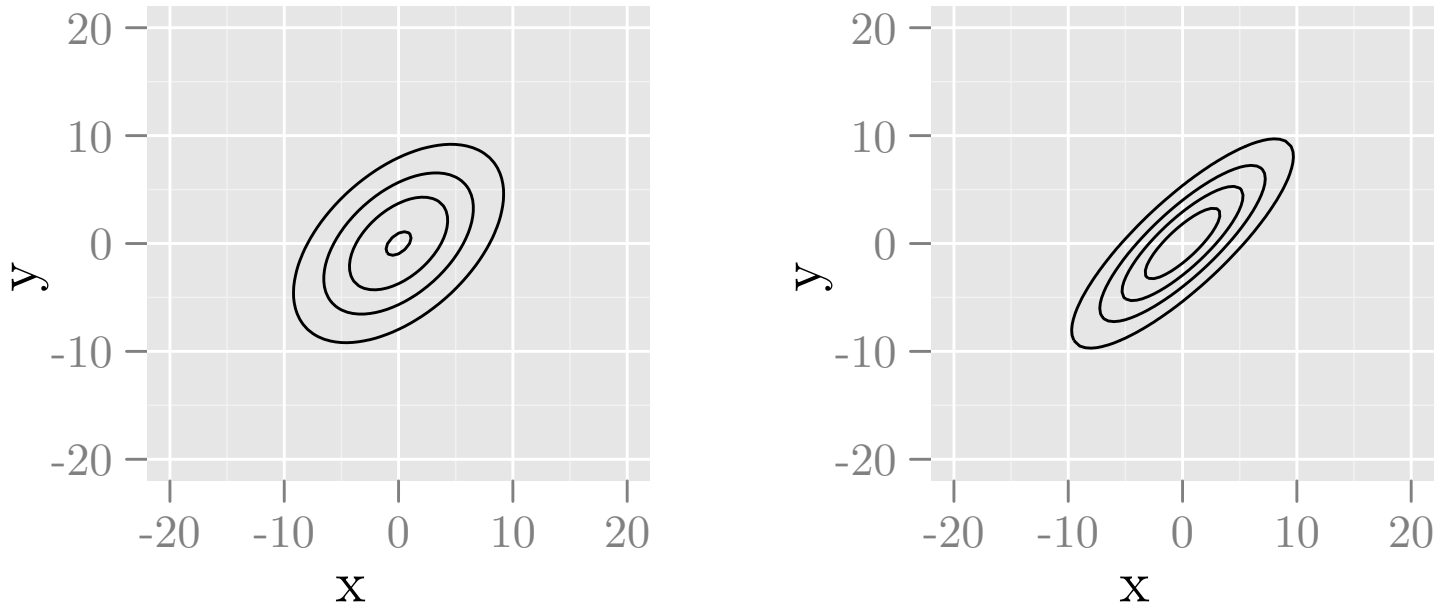
$$\begin{aligned} f_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{(\sigma_{11})\cdots(\sigma_{pp})}} \\ &\quad \times \exp\left\{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_{11}} - \cdots - \frac{1}{2}\frac{(x_p-\mu_p)^2}{\sigma_{pp}}\right\} \\ &= f_{x_1}(x_1)\cdot f_{x_2}(x_2)\cdot \cdots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

- Shape of the MVN density



(Fig 4.2 from RC)

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



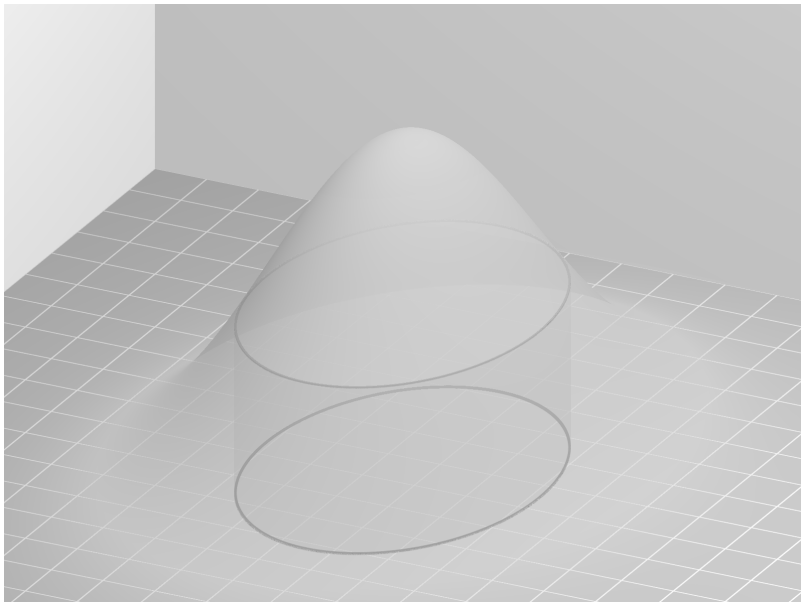
(Fig 4.3 from RC)

$\sigma_{11} = \sigma_{22}$ 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 $\underset{q\times p}{\mathbf{A}}$

   - $\mathbf{Ax} + \mathbf{c} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$
   - $\mathbf{c}'\mathbf{x} \sim N_1(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$
   - $E\{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})\} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu} - \boldsymbol{\mu}) = \mathbf{0_p}$
     $\text{var}\{\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})\} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I_p}$
     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

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

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

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

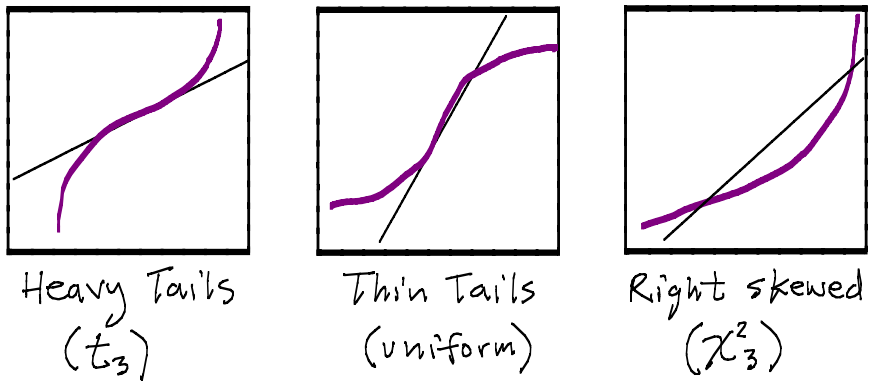
- Though normality of univariate & bivariate subsets of $\underset{p\times 1}{\mathbf{x}}$ does not guarantee MVN'ity, in practice, 1-D and 2-D investigations are often sufficient

  1-D Tools

  - 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

<u>2-D Tools</u>

- 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 $\beta(\frac{p}{2}, \frac{1}{2}(n-p-1))$

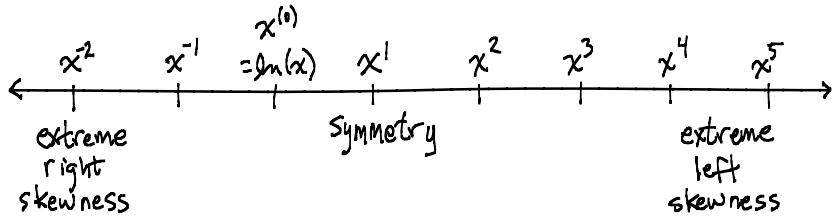
    y-axis: $\frac{n}{(n-1)^2} D^2_{(i)}$

– "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

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

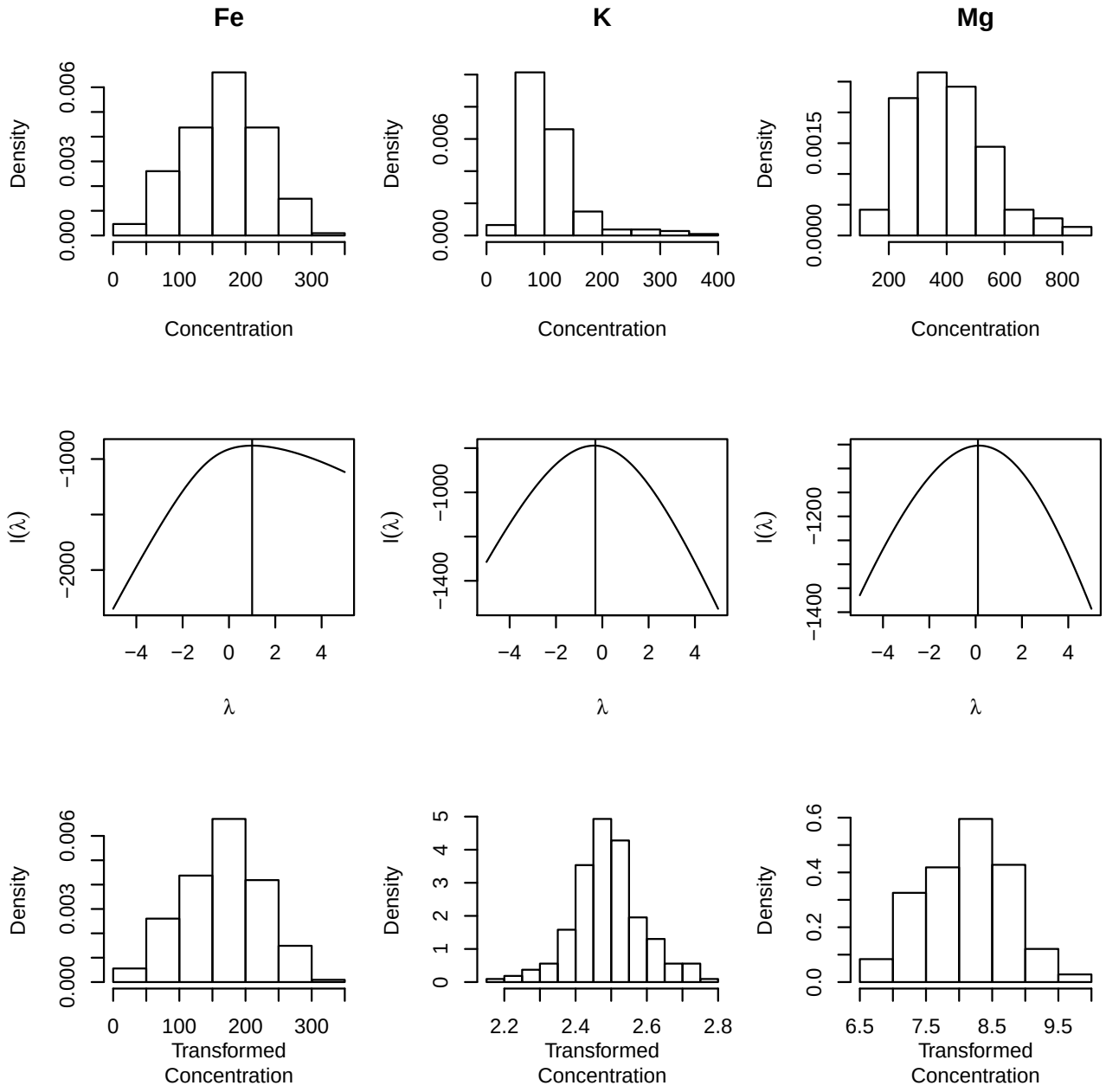
is the maximum likelihood estimate of the variance of $x^{(\lambda)}$ and $\overline{x^{(\lambda)}}$ is the sample mean of the n transformed observations

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
Concentration
l(λ)
λ
Transformed
Concentration

G Maximum Likelihood

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

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

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

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

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

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

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

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

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

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

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

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

So,

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

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

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

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 $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$

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

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 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 $\mathbf{x}$ 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

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

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

Missing at random??

Two steps of Algorithm:

1. Expectation (Prediction) Step

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

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

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

   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

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

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

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

where

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

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.