LECTURE NOTES: STATISTICS 537 CLASSICAL MULTIVARIATE ANALYSIS SPRING 2004

Robert J. Boik
Department of Mathematical Sciences
Montana State University — Bozeman

April 28, 2004

Contents

Chapter 1

REFERENCE SOURCES

1.1 COURSE SYLLABUS

- Required Texts
  - Rencher, A. C. (1998). *Multivariate Statistical Inference and Applications*. New York: Wiley.
- Instructor
  - Robert J. Boik, 2–260 Wilson, 994-5339, Rjboik@math.montana.edu.
  - Office Hours: Monday 11:00–11:50; Tuesday 2:10–3:00; Wednesday 11:00–11:50; Thursday 2:10–3:00.
- Course Home Page: <http://www.math.montana.edu/∼rjboik/classes/537/stat.537.html>
- Holidays & Other "No Class" Days: Monday Jan 19 (Martin Luther King), Monday January 26: U. Co. talk, Monday Feb 16 (Presidents Day), Wednesday March 3 (Exam Exchange Day), Monday–Friday Mar 15–19 (Spring Break), Friday April 9 (University Day).
- HW: Discussion about HW problems with colleagues is allowed, but written work must be done independently. Late HW will not be accepted without prior arrangements.
- Grading: A Midterm exam will be given on Wednesday March 3 at 6:00-8:00 PM (20%) in 1-153 Wilson. A Final exam will be given on Tuesday May 4 at 8:00–9:50 AM (40%) in 1-153 Wilson. The remaining 40% is from HW.

Syllabus

1. Introduction: Univariate versus Multivariate Analysis
2. Multivariate Data & Multivariate Distributions (Ch. 1, 2)
   (a) Expectation and Dispersion of Random Matrices
   (b) Multivariate Normal: Conditional and Marginal
      i. Detecting Departure from MVN
      ii. Transformations of Multivariate Data
   (c) Correlation, Partial Correlation, and Regression
   (d) Wishart and Conditional Wishart Distributions
   (e) Maximum Likelihood Estimation from MVN
      i. Complete Data
      ii. Incomplete Data: the EM Algorithm
   (f) Robust Estimation
3. Multivariate Linear Models (Ch. 3, 4, 7)

   (a) Hotelling's T^2 Tests
   (b) The General Linear Model
   (c) Test Statistics & Simultaneous Inference
   (d) Classic Analysis of Repeated Measures & Growth Curves
   (e) Generalized Analysis of Repeated Measures & Longitudinal Data
   (f) Introduction to Proc Mixed
4. Selected Inferences on Covariance Matrices (Ch. 8)
   (a) Tests of Sphericity
   (b) Tests of Homogeneity
   (c) Tests of Independence
   (d) Canonical Correlation
5. Discriminant & Classification Analysis (Ch. 5, 6)
6. Principal Components (Ch. 9)
   (a) Common Principal Components
   (b) Principal Components of Correlation Matrices
7. Factor Analysis (Ch. 10)
8. Cluster Analysis
9. Classification Trees

1.2 REFERENCE BOOKS

1.2.1 Introductory Books

1. Afifi, A. A., & Clark, V. (1990). *Computer-Aided Multivariate Analysis*. (Second Edition), New York: Van Nostrand Reinhold.
2. Dillon, W. R., & Goldstein, M. (1984). *Multivariate Analysis—Methods and Applications*. New York: Wiley.
3. Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis with Readings*, 4^{th} edition, Englewood Cliffs, NJ: Prentice Hall.
4. Johnson, R. A., & Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis*. (Fourth Edition), Englewood Cliffs, NJ: Prentice Hall.
5. Johnson, D. E. (1998). *Applied Multivariate Methods for Data Analysts*, Pacific Grove, CA: Duxbury Press.
6. Manly, B. F. J. (1994). *Multivariate Statistical Methods: A Primer*, London: Chapman & Hall.

1.2.2 Intermediate Books

1. Rencher, A. C. (2002). *Methods of Multivariate Analysis*, Second Edition, New York: Wiley.
2. Harris, R. J. (1985). *A Primer of Multivariate Statistics*. (Second Edition), Orlando, Florida: Academic Press.
3. Morrison, D. F. (1990). *Multivariate Statistical Methods*. (Third Edition), New York: McGraw Hill.
4. Seber, G. A. F. (1984). *Multivariate Observations*. New York: Wiley
5. Timm, N. H. (1975). *Multivariate Analysis with Applications in Education and Psychology*. Belmont, CA: Wadsworth Publishing Company.

1.2.3 Advanced Books

1. Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. (Third Edition), New York: Wiley.

2. Bilodeau, M. & Bremer, D. (1999). *Theory of Multivariate Statistics*, New York: Springer-Verlag.

3. Flury, B. (1988). *Common Principal Components & Related Multivariate Models*. New York: Wiley.

4. Giri, N. C. (2003). *Multivariate Statistical Analysis*. New York: Marcel Dekker, Inc.

5. Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. New York: Academic Press.

6. Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

7. Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. (Second Edition), Malabar, FL: Krieger Publishing Company.

8. Siotani, M., Hayakawa, T., & Fujikoshi, Y. (1985). *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*. Columbus, Ohio: American Sciences Press.

9. Srivastava, M. S., & Khatri, C. G. (1979). *An Introduction to Multivariate Statistics*. New York: Elsevier North Holland Inc.

Chapter 2

RANDOM VARIABLES, VECTORS, & MATRICES

2.1 EXPECTATION AND COVARIANCE OPERATORS

2.1.1 Expectation of Random Matrices

Let y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$ be a collection of random variables. Then $\mathbf{Y} = \{y_{ij}\}$ is a random matrix. Suppose that $\mathrm{E}(y_{ij}) = \mu_{ij} < \infty$. Let $\mathbf{M} = \{\mu_{ij}\}$. Then, the following can be established.

1. $\mathrm{E}(\mathbf{Y}) = \{\mathrm{E}(y_{ij})\} = \mathbf{M}$.

2. $\mathrm{E}(\mathbf{AYB} + \mathbf{C}) = \mathbf{A}\mathrm{E}(\mathbf{Y})\mathbf{B} + \mathbf{C} = \mathbf{AMB} + \mathbf{C}$ where $\mathbf{A}$: $p \times n$, $\mathbf{B}$: $d \times r$, and $\mathbf{C}$: $p \times r$ are matrices of constants.

2.1.2 Variance and Covariance of Random Vectors

Let $\mathbf{y}$ be an $n \times 1$ random vector and let $\mathbf{x}$ be an $r \times 1$ random vector.

1. $\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathrm{E}\{[\mathbf{x} - \mathrm{E}(\mathbf{x})][\mathbf{y} - \mathrm{E}(\mathbf{y})]'\} = \mathrm{E}(\mathbf{x}\mathbf{y}') - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{y}')$.

2. $\mathrm{Cov}(\mathbf{Ax}, \mathbf{By}) = \mathrm{E}\{[\mathbf{Ax} - \mathbf{A}\mathrm{E}(\mathbf{x})][\mathbf{By} - \mathbf{B}\mathrm{E}(\mathbf{y})]'\}$

$$= \mathrm{E}\{\mathbf{A}[\mathbf{x} - \mathrm{E}(\mathbf{x})][\mathbf{y} - \mathrm{E}(\mathbf{y})]'\mathbf{B}'\} = \mathbf{A}\,\mathrm{Cov}(\mathbf{x}, \mathbf{y})\mathbf{B}'.$$

3. $\mathrm{Var}(\mathbf{y}) = \mathrm{Cov}(\mathbf{y}, \mathbf{y})$.

4. Using (2) and (3), $\mathrm{Var}(\mathbf{Ay}) = \mathrm{Cov}(\mathbf{Ay}, \mathbf{Ay}) = \mathbf{A}\,\mathrm{Var}(\mathbf{y})\mathbf{A}'$.

2.1.3 Dispersion of Random Matrices

2.1.4 *General Setting*

Let $\mathbf{Y}$: $n \times d$ be a random matrix. Partition $\mathbf{Y}$ as $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_d \end{pmatrix}$. Denote $\mathrm{Var}(\mathbf{y}_i)$ by $\mathbf{\Sigma}_{ii}$ and denote $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)$ by $\mathbf{\Sigma}_{ij}$ for $i = 1, \ldots, d$, $j = 1, \ldots, d$. Note that each $\mathbf{\Sigma}_{ij}$ is $n \times n$. Then,

$$\mathrm{Disp}(\mathbf{Y}) \stackrel{\mathrm{def}}{=} \mathrm{Var}[\mathrm{Vec}(\mathbf{Y})] = \{\mathbf{\Sigma}_{ij}\}\text{: } nd \times nd.$$

2.1.5 *Standard Setting*

Let $\mathbf{Y}$ be an $n \times d$ matrix. Partition $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}.$$

In the standard multivariate setting, the rows of $\mathbf{Y}$ represent a random sample from $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It follows that $\mathrm{E}(\mathbf{y}_i) = \boldsymbol{\mu}\ \forall\ i$, $\mathrm{Var}(\mathbf{y}_i) = \boldsymbol{\Sigma}\ \forall\ i$, and $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{0}$ for $i \neq j$. The following results are readily established:

1. $\mathrm{E}(\mathbf{Y}) = \mathbf{M} = \mathbf{1}_n\boldsymbol{\mu}'$,
2. $\mathrm{Disp}(\mathbf{Y}) = (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)$, and
3. $\mathrm{vec}(\mathbf{Y}) \sim [(\mathbf{I}_d \otimes \mathbf{1}_n)\boldsymbol{\mu}, (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)]$.

2.1.6 Expectation of Univariate Quadratic Forms

Theorem 2.1 *Let* $\mathbf{y}$ *be an* $n \times 1$ *random vector with mean* $\boldsymbol{\mu}$ *and variance* $\boldsymbol{\Sigma}$*. Let* $\mathbf{T}\colon n \times n$ *be a matrix of constants. Then,*

$$\mathrm{E}(\mathbf{y}'\mathbf{T}\mathbf{y}) = \mathrm{tr}(\mathbf{T}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu}.$$

Proof: $\mathrm{E}(\mathbf{y}'\mathbf{T}\mathbf{y}) = \mathrm{E}[\mathrm{tr}(\mathbf{y}'\mathbf{T}\mathbf{y})] = \mathrm{E}[\mathrm{tr}(\mathbf{T}\mathbf{y}\mathbf{y}')]$

$$= \mathrm{tr}[\mathrm{E}(\mathbf{T}\mathbf{y}\mathbf{y}')] = \mathrm{tr}[\mathbf{T}\mathrm{E}(\mathbf{y}\mathbf{y}')] = \mathrm{tr}[\mathbf{T}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')].$$

□

2.1.7 Expectation of Multivariate Quadratic Forms

Theorem 2.2 *Let* $\mathbf{Y}$ *be an* $n \times d$ *matrix with distribution* $\mathrm{vec}(\mathbf{Y}) \sim [\mathrm{vec}(\mathbf{M}), \boldsymbol{\Sigma}]$*. Partition* $\mathbf{Y}$ *and* $\mathbf{M}$ *as*

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_d \end{pmatrix} \text{ and } \mathbf{M} = \begin{pmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \cdots & \boldsymbol{\mu}_d \end{pmatrix}.$$

Denote $\mathrm{Var}(\mathbf{y}_i)$ *by* $\boldsymbol{\Sigma}_{ii}$*, and denote* $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)$ *by* $\boldsymbol{\Sigma}_{ij}$*. Note that* $\boldsymbol{\Sigma}_{ij}$ *is* $n \times n$ *and that* $\boldsymbol{\Sigma}_{ij}' = \boldsymbol{\Sigma}_{ji}$*. Let* $\mathbf{T}\colon n \times n$ *be a matrix of constants. Then,*

$$\mathrm{E}(\mathbf{Y}'\mathbf{T}\mathbf{Y}) = \mathbf{M}'\mathbf{T}\mathbf{M} + T_d[(\mathbf{I}_d \otimes \mathbf{T}')\boldsymbol{\Sigma}],$$

where $T_d(\ \cdot\ )$ *is the generalized trace operator. See the STAT 505 notes for a description of this operator. Furthermore, if* $\mathbf{T}$ *is symmetric, then*

$$\mathrm{E}(\mathbf{Y}'\mathbf{T}\mathbf{Y}) = \mathbf{M}'\mathbf{T}\mathbf{M} + T_d[(\mathbf{I}_d \otimes \mathbf{T})\boldsymbol{\Sigma}].$$

Proof: Write $\mathbf{Y}'\mathbf{T}\mathbf{Y}$ *as*

$$\mathbf{Y}'\mathbf{T}\mathbf{Y} = \{\mathbf{y}_i'\mathbf{T}\mathbf{y}_j\},$$

and use

$$\begin{aligned}
\mathrm{E}(\mathbf{y}_i'\mathbf{T}\mathbf{y}_j) &= \mathrm{E}\left[\mathrm{tr}\left(\mathbf{T}\mathbf{y}_j\mathbf{y}_i'\right)\right] \\
&= \mathrm{tr}\left[\mathrm{E}\left(\mathbf{T}\mathbf{y}_j\mathbf{y}_i'\right)\right] \\
&= \mathrm{tr}\left[\mathbf{T}(\boldsymbol{\Sigma}_{ji} + \boldsymbol{\mu}_j\boldsymbol{\mu}_i')\right] \\
&= \mathrm{tr}\left(\mathbf{T}\boldsymbol{\Sigma}_{ji}\right) + \boldsymbol{\mu}_i'\mathbf{T}\boldsymbol{\mu}_j \\
&= \mathrm{tr}\left(\boldsymbol{\Sigma}_{ij}\mathbf{T}'\right) + \boldsymbol{\mu}_i'\mathbf{T}\boldsymbol{\mu}_j \ \textit{because}\ \mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}') \\
&= \mathrm{tr}\left(\mathbf{T}'\boldsymbol{\Sigma}_{ij}\right) + \boldsymbol{\mu}_i'\mathbf{T}\boldsymbol{\mu}_j \ \textit{because}\ \mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}).
\end{aligned}$$

□

For the following corollaries, assume that $\mathbf{T}$ is symmetric.

Corollary 1: If $\mathbf{\Sigma} = (\mathbf{\Sigma}_d \otimes \mathbf{\Omega})$, then $\mathrm{E}(\mathbf{Y}'\mathbf{T}\mathbf{Y}) = \mathbf{M}'\mathbf{T}\mathbf{M} + \mathrm{trace}(\mathbf{T}\mathbf{\Omega})\,\mathbf{\Sigma}$.

Corollary 2: If $\mathbf{\Sigma} = (\mathbf{\Sigma}_d \otimes \mathbf{I}_n)$, then $\mathrm{E}(\mathbf{Y}'\mathbf{T}\mathbf{Y}) = \mathbf{M}'\mathbf{T}\mathbf{M} + \mathrm{trace}(\mathbf{T})\,\mathbf{\Sigma}$.

Corollary 3: If $\mathbf{\Sigma} = (\mathbf{\Sigma}_d \otimes \mathbf{I}_n)$ and $\mathbf{M} = \mathbf{1}_n\boldsymbol{\mu}'$, then $\mathrm{E}(\mathbf{Y}'\mathbf{T}\mathbf{Y}) = \boldsymbol{\mu}\boldsymbol{\mu}'(\mathbf{1}_n'\mathbf{T}\mathbf{1}_n) + \mathrm{trace}(\mathbf{T})\,\mathbf{\Sigma}_d$.

Corollary 4: If $\mathbf{\Sigma} = (\mathbf{\Sigma}_d \otimes \mathbf{I}_n)$ and $\mathbf{M} = \mathbf{X}\mathbf{B}$, then $\mathrm{E}(\mathbf{Y}'\mathbf{T}\mathbf{Y}) = \mathbf{B}'\mathbf{X}'\mathbf{T}\mathbf{X}\mathbf{B}' + \mathrm{trace}(\mathbf{T})\,\mathbf{\Sigma}_d$.

2.1.8 Sample Means and Variances

Let $\mathbf{Y}$ be an $n \times d$ matrix that follows the standard multivariate setting. Let $\overline{\mathbf{y}} = n^{-1}\mathbf{Y}'\mathbf{1}_n$ and $\mathbf{S} = (n-1)^{-1}\mathbf{Y}'[\mathbf{I} - n^{-1}\mathbf{1}_n\mathbf{1}_n']\mathbf{Y}$. Then

1. $\mathrm{E}(\overline{\mathbf{y}}) = \boldsymbol{\mu}$,
2. The BLUE of $\boldsymbol{\mu}$ is $\overline{\mathbf{y}}$,
3. $\mathrm{Var}(\overline{\mathbf{y}}) = n^{-1}\mathbf{\Sigma}$, and
4. $\mathrm{E}(\mathbf{S}) = \mathbf{\Sigma}$.

2.1.9 Regression Coefficients and Variances

Extension of Standard Multivariate Setup: Let $\mathbf{Y}$ be an $n \times d$ matrix with expectation $\mathbf{M}$. Partition $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}.$$

Suppose that $\mathrm{Var}(\mathbf{y}_i) = \mathbf{\Sigma}\ \forall\ i$, $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{0}$ for $i \neq j$, and $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\mathbf{B}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants having rank-p and $\mathbf{B}$ is a $p \times d$ matrix of regression coefficients. That is, $\mathrm{Disp}(\mathbf{Y}) = (\mathbf{\Sigma} \otimes \mathbf{I}_n)$ and $\mathrm{vec}(\mathbf{Y}) \sim [(\mathbf{I}_d \otimes \mathbf{X}_n)\boldsymbol{\beta}, (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$, where $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$. Let $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and $\mathbf{S} = (n-p)^{-1}\mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}$, where $\mathbf{H} = \mathrm{ppo}(\mathbf{X}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then

1. $\mathrm{E}(\widehat{\mathbf{B}}) = \mathbf{B}$,
2. The BLUE of $\mathbf{B}$ is $\widehat{\mathbf{B}}$,
3. $\mathrm{Disp}(\widehat{\mathbf{B}})) = \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}$, and
4. $\mathrm{E}(\mathbf{S}) = \mathbf{\Sigma}$.

2.2 MULTIVARIATE NORMAL DISTRIBUTION

Suppose $\mathbf{y}\colon n \times 1$ is a random vector with joint probability density function

$$f(\mathbf{y}) = \frac{\exp\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\}}{|\mathbf{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}}$$

for $\mathbf{\Sigma} > 0$, $\boldsymbol{\mu} \in R^n$, and $\mathbf{y} \in R^n$. Then, $\mathbf{y}$ is said to have a multivariate normal distribution. The notations $\mathbf{y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \mathbf{\Sigma})$ and $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ are used to indicate that the elements of the random vector $\mathbf{y}$ are jointly distributed as an n dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$. If $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$, then the pdf simplifies to

$$f(\mathbf{y}) = \frac{\exp\{-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})'(\mathbf{y}-\boldsymbol{\mu})\}}{(2\pi\sigma^2)^{\frac{n}{2}}} = \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2\}}{(2\pi\sigma^2)^{\frac{n}{2}}}$$

Caution: the notation $\mathbf{y} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used to indicate that the elements of the random vector $\mathbf{y}$ are jointly distributed such that $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{y}) = \boldsymbol{\Sigma}$. Obviously, $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \mathbf{y} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ but $\mathbf{y} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \not\Rightarrow \mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If $\boldsymbol{\Sigma}$ is not positive definite (i.e., it is positive semi-definite), then $\mathbf{y}$ is said to have a singular normal distribution: $\mathbf{y} \sim \mathrm{SN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this case, $\mathbf{y}$ does not have a density function, but $\mathbf{y}$ still has a distribution function.

2.2.1 Multivariate Normal Matrices

Suppose $\mathbf{Y}\colon n \times d$ is a random matrix. Let $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$. If the elements of $\mathbf{y}$ have joint density function

$$f(\mathbf{y}) = \frac{\exp\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{nd}{2}}},$$

where $\boldsymbol{\Sigma}$ is an $nd \times nd$ positive definite matrix, $\boldsymbol{\mu} \in R^{nd}$, and $\mathbf{y} \in R^{nd}$, then $\mathbf{Y}$ is said to have a multivariate normal distribution. The notations $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}_{nd}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{M}), \boldsymbol{\Sigma}]$, where $\boldsymbol{\mu} = \mathrm{vec}(\mathbf{M})$, are used to indicate that the elements of the random matrix $\mathbf{Y}$ are jointly distributed as an nd dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

If $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_d \otimes \boldsymbol{\Omega}$, then the density function can be written as

$$f(\mathbf{Y}) = \frac{\exp\left\{-\frac{1}{2}\,\mathrm{trace}\left[(\mathbf{Y}-\mathbf{M})\boldsymbol{\Sigma}_d^{-1}(\mathbf{Y}-\mathbf{M})'\boldsymbol{\Omega}^{-1}\right]\right\}}{|\boldsymbol{\Omega}|^{\frac{d}{2}}|\boldsymbol{\Sigma}_d|^{\frac{n}{2}}(2\pi)^{\frac{nd}{2}}}.$$

In the standard multivariate setup, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_d \otimes \mathbf{I}_n$ and the density simplifies to

$$f(\mathbf{Y}) = \frac{\exp\left\{-\frac{1}{2}\,\mathrm{trace}\left[(\mathbf{Y}-\mathbf{M})'(\mathbf{Y}-\mathbf{M})\boldsymbol{\Sigma}_d^{-1}\right]\right\}}{|\boldsymbol{\Sigma}_d|^{\frac{n}{2}}(2\pi)^{\frac{nd}{2}}}.$$

2.2.2 Properties of the MVN Distribution

1. $\int_{-\infty}^{\infty} f(\mathbf{y})\, d\mathbf{y} = 1$
2. Moment generating function (MGF): $M_{\mathbf{y}}(t) = \mathrm{E}[\exp(\mathbf{t}'\mathbf{y})] = \exp[\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})]$. Proof: in class.
3. $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$. Proof: use MGF.
4. $\mathrm{Var}(\mathbf{y}) = \boldsymbol{\Sigma}$. proof: use MGF.
5. $\mathbf{A}\mathbf{y} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ for any matrix of constants: $\mathbf{A}\colon r \times n$. Proof: use MGF. Note that if $\mathbf{A}$ does not have full row rank, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ is singular and $\mathbf{A}\mathbf{y} \sim \mathrm{SN}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. If $\mathbf{A}$ is a random matrix, then $\mathbf{A}\mathbf{y}$ may or may not have a multivariate normal distribution.
6. If $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{M}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}]$, then

$$M_{\mathbf{Y}}(\mathbf{T}) = \mathrm{E}\left[\exp[\mathrm{trace}(\mathbf{T}'\mathbf{Y})]\right] = \exp\left[\mathrm{trace}(\mathbf{T}'\mathbf{M}) + \frac{1}{2}\,\mathrm{trace}(\mathbf{T}'\boldsymbol{\Omega}\mathbf{T}\boldsymbol{\Sigma})\right].$$

7. If $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{M}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}]$, then $\mathrm{vec}(\mathbf{AYG}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{AMG}), \mathbf{G}'\boldsymbol{\Sigma}\mathbf{G} \otimes \mathbf{A}\boldsymbol{\Omega}\mathbf{A}']$ for constant matrices $\mathbf{A}$ and $\mathbf{G}$. If $\mathbf{A}$ does not have full row rank or $\mathbf{G}$ does not have full column rank, then the distribution is singular and the density does not exist.
8. In many applications, the vector of means can be written as $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an $n \times p$ matrix with rank-r.
9. In many applications, the matrix of means can be written as $\mathbf{M} = \mathbf{X}\mathbf{B}$, where $\mathbf{X}$ is an $n \times p$ matrix with rank-r. Note, $\boldsymbol{\mu} = \mathrm{vec}(\mathbf{M}) = (\mathbf{I}_d \otimes \mathbf{X})\boldsymbol{\beta}$, where $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$.

2.3 CONDITIONAL MULTIVARIATE NORMAL DISTRIBUTIONS

Let $\mathbf{y}$ be a random vector and denote a realization of the random vector by $\ddot{\mathbf{y}}$. Suppose that $\mathbf{y}\colon n \times 1$ is distributed as $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Partition $\mathbf{y}$ as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$$

where $\mathbf{y}_1$ is $p \times 1$ and $\mathbf{y}_2$ is $(n-p) \times 1$. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, conformably, as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and

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

Theorem 2.3 (Conditional Multivariate Normal) *Conditional on $\mathbf{y}_2 = \ddot{\mathbf{y}}_2$, $\mathbf{y}_1$ still has a joint normal distribution. In particular,*

$$\mathbf{y}_1 | (\mathbf{y}_2 = \ddot{\mathbf{y}}_2) \sim \mathrm{N}(\boldsymbol{\mu}_{1\cdot 2}, \boldsymbol{\Sigma}_{11\cdot 2}),$$

where $\boldsymbol{\mu}_{1\cdot 2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\ddot{\mathbf{y}}_2 - \boldsymbol{\mu}_2)$ and $\boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. If $\boldsymbol{\Sigma}_{21} \in \mathcal{R}(\boldsymbol{\Sigma}_{22})$, then $\boldsymbol{\Sigma}_{22}$ need not be nonsingular. Simply replace $\boldsymbol{\Sigma}_{22}^{-1}$ by $\boldsymbol{\Sigma}_{22}^{-}$.

Proof: In class. □

Let $\mathbf{Y}$ be a $n \times d$ matrix with distribution $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{M}), \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}]$. Partition $\mathbf{Y}$ as $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 \end{pmatrix}$, where $\mathbf{Y}_1$ is $n \times d_1$, $\mathbf{Y}_2$ is $n \times d_2$, and $d_1 + d_2 = d$. Partition $\mathbf{M}$ and $\boldsymbol{\Sigma}$ conformably as $\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{M}_2 \end{pmatrix}$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \otimes \boldsymbol{\Omega}.$$

It follows from the previous result that, conditional on $\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2$, $\mathbf{Y}_1$ still has a joint normal distribution. In particular,

$$\mathrm{vec}(\mathbf{Y}_1) | (\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2) \sim \mathrm{N}[\mathrm{vec}(\mathbf{M}_{1\cdot 2}), \boldsymbol{\Sigma}_{11\cdot 2}],$$

where

$$\mathbf{M}_{1\cdot 2} = \mathbf{M}_1 + (\ddot{\mathbf{Y}}_2 - \mathbf{M}_2)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

and

$$\boldsymbol{\Sigma}_{11\cdot 2} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \otimes \boldsymbol{\Omega}.$$

If $\boldsymbol{\Sigma}_{21} \in \mathcal{R}(\boldsymbol{\Sigma}_{22})$, then $\boldsymbol{\Sigma}_{22}$ need not be nonsingular. Simply replace $\boldsymbol{\Sigma}_{22}^{-1}$ by $\boldsymbol{\Sigma}_{22}^{-}$.

2.3.1 Regression Application

Let $\mathbf{y}$ be a random d-vector and let $\mathbf{x}$ be a random p-vector. Suppose that

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right].$$

Then, the distribution of $\mathbf{y}$, conditional on $\mathbf{x} = \ddot{\mathbf{x}}$, is

$$\mathbf{y} | (\mathbf{x} = \ddot{\mathbf{x}}) \sim \mathrm{N}(\boldsymbol{\beta}_0 + \mathbf{B}_1'\ddot{\mathbf{x}}, \boldsymbol{\Sigma}_{yy\cdot x}),$$

where $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_y - \mathbf{B}_1'\boldsymbol{\mu}_x$; $\mathbf{B}_1 = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$; and $\boldsymbol{\Sigma}_{yy\cdot x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$.

If we have a random sample $\begin{pmatrix} \mathbf{y}_i' & \mathbf{x}_i' \end{pmatrix}'$ for $i = 1, \ldots, n$, then the distribution of

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} \text{ conditional on } \mathbf{X} = \ddot{\mathbf{X}} = \begin{pmatrix} \ddot{\mathbf{x}}_1' \\ \ddot{\mathbf{x}}_2' \\ \vdots \\ \ddot{\mathbf{x}}_n' \end{pmatrix}$$

is

$$\text{vec}(\mathbf{Y}) \sim \text{N}\left[\text{vec}(\mathbf{X}^*\mathbf{B}), \boldsymbol{\Sigma}_{yy\cdot x} \otimes \mathbf{I}_n\right],$$

where

$$\mathbf{X}^* = \begin{pmatrix}\mathbf{1}_n & \ddot{\mathbf{X}}\end{pmatrix}; \quad \mathbf{B} = \begin{pmatrix}\boldsymbol{\beta}_0' \\ \mathbf{B}_1\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_y' - \boldsymbol{\mu}_x'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\end{pmatrix};$$

and $\boldsymbol{\Sigma}_{yy\cdot x}$ is defined above. Conditional on $\mathbf{X} = \ddot{\mathbf{X}}$, unbiased estimators of $\mathbf{B}$ and $\boldsymbol{\Sigma}_{yy\cdot x}$ are given by

$$\widehat{\mathbf{B}} = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y} \text{ and } \widehat{\boldsymbol{\Sigma}}_{xx\cdot y} = \frac{\mathbf{Y}'\left[\mathbf{I}_n - \text{ppo}\left(\mathbf{X}^*\right)\right]\mathbf{Y}}{n-p-1}.$$

2.3.2 Some Other Conditional Results

Suppose that $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are random matrices.

1. If the distribution of $\mathbf{X}|(\mathbf{Y} = \ddot{\mathbf{Y}})$ does not depend on $\ddot{\mathbf{Y}}$, it follows that $\mathbf{X}$ and $\mathbf{Y}$ are independent. Of course, it works the other way too: if $\mathbf{X}$ and $\mathbf{Y}$ are independent, then the distribution of $\mathbf{X}|(\mathbf{Y} = \ddot{\mathbf{Y}})$ does not depend on $\ddot{\mathbf{Y}}$.

2. Suppose that $\mathbf{X}$ and $\mathbf{Z}$ are independent, conditional on $\mathbf{Y} = \ddot{\mathbf{Y}}$. That is, $\mathbf{X}$ and $\mathbf{Z}$ are conditionally independent: $(\mathbf{X} \perp\!\!\!\perp \mathbf{Z})|(\mathbf{Y} = \ddot{\mathbf{Y}})$. Suppose, also, that $\mathbf{X}$ and $\mathbf{Y}$ are independent (unconditionally). Then $\mathbf{X}$ and $\mathbf{Z}$ are unconditionally independent.

2.4 DETECTING DEPARTURE FROM NORMALITY

2.4.1 Univariate Procedures

Assume that $Y_1, Y_2, \ldots, Y_n$ is a random sample from a distribution with cdf $F_Y(y)$.

2.4.1.1 *QQ Plots*

The letters QQ stand for quantile-quantile. The sample or empirical quantiles are equal to the $100\frac{1}{n}, 100\frac{2}{n}, \ldots, 100\frac{n}{n}$ sample percentiles. The $100\frac{i}{n}$ sample percentile, in turn, is equal to $Y_{(i)}$, the i^{th} order statistic. If $F_Y(y) = F(y)$, then the $100\frac{i}{n}$ theoretical quantile can be defined as $F_Y^{-1}(\alpha_i)$, where $\alpha_i = (i - 3/8)/(n + 1/4)$. The reason that α_i is not defined as i/n is that in many distributions the 100^{th} percentile is ∞.

A QQ plot is a plot of the empirical quantiles (Y axis) against the theoretical quantiles (X axis). The table below displays the pairs to be plotted.

i	α_i	X Axis	Y Axis
1	$(1-\frac{3}{8})/(n+\frac{1}{4})$	$F^{-1}(\alpha_1)$	$Y_{(1)}$
2	$(2-\frac{3}{8})/(n+\frac{1}{4})$	$F^{-1}(\alpha_2)$	$Y_{(2)}$
3	$(3-\frac{3}{8})/(n+\frac{1}{4})$	$F^{-1}(\alpha_3)$	$Y_{(3)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$(n-\frac{3}{8})/(n+\frac{1}{4})$	$F^{-1}(\alpha_n)$	$Y_{(n)}$

The points in the plot should lie in a straight line. If they do not (within sampling error), then there is evidence that the data do not come from distribution $F(y)$.

To construct a normal QQ plot (usually called a normal plot), first standardize the data to have mean zero and standard deviation one: $Z_i = (Y_i - \overline{Y})/S$. Then plot $Z_{(i)}$ against $\Phi^{-1}(\alpha_i)$, where Φ is the cdf of $\text{N}(0,1)$. Normal plots can be constructed in SAS using proc capability.

2.4.1.2 *Shapiro-Wilk*

The Shapiro-Wilk statistic is obtained by regressing the order statistics on the theoretical quantiles from the normal distribution. Ordinary regression is not appropriate because the order statistics are not independent of one another. Generalized least squares is used. The Shapiro-Wilk regression coefficient is bounded above by one. The null hypothesis of normality is rejected if the regression coefficient is small. SAS uses a normalizing transformation on the Shapiro-Wilk test statistic to obtain a p-value.

2.4.1.3 *Coefficients of Skewness and Kurtosis*

The univariate skewness and kurtosis coefficients are

$$\kappa_3 = \frac{\mathrm{E}(Y-\mu)^3}{\sigma^3} \text{ and } \kappa_4 = \frac{\mathrm{E}(Y-\mu)^4}{\sigma^4} - 3,$$

where $\mu = \mathrm{E}(Y)$ and $\sigma^2 = \mathrm{var}(Y)$. Under normality $\kappa_3 = 0$ and $\kappa_4 = 0$. Sample estimators of κ_3 and κ_4 are

$$\begin{aligned}
\widehat{\kappa}_3 &= \frac{k_3}{S^3} \text{ and } \widehat{\kappa}_4 = \frac{k_4}{S^4}, \text{ where} \\
S^2 &= \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}, \\
k_3 &= \frac{n\sum_{i=1}^{n}(Y_i - \overline{Y})^3}{(n-1)(n-2)}, \text{ and} \\
k_4 &= \frac{n(n+1)\sum_{i=1}^{n}(Y_i - \overline{Y})^4 - 3(n-1)\left\{\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right\}^2}{(n-1)(n-2)(n-3)}.
\end{aligned}$$

The statistics k_3 and k_4 are unbiased estimators of the 3^{rd} and 4^{th} cumulants:

$$\mathrm{E}(k_3) = \mathrm{E}(Y-\mu)^3 \text{ and } \mathrm{E}(k_4) = \mathrm{E}(Y-\mu)^4 - 3\sigma^4.$$

It can be shown (A. Stuart & K. Ord, *Kendall's Advanced Theory of Statistics*, Vol 1, 5^{th} ed., Oxford University Press, 1987) that if the distribution of Y is normal, then $\widehat{\kappa}_3$ and $\widehat{\kappa}_4$ are distributed approximately normal (in large samples) with means zero and variances

$$\mathrm{Var}(\widehat{\kappa}_3) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \text{ and } \mathrm{Var}(\widehat{\kappa}_4) = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}.$$

Tables A1 and A3 in Rencher (2002) give percentiles of biased estimators of κ_3 and κ_4.

2.4.1.4 *Kolmogorov-Smirnov*

Consider a test of $\mathrm{H}_0\colon F_Y(y) = F_0(y)$ against the alternative $\mathrm{H}_a\colon F_Y(y) \neq F_0(y)$. The Kolmogorov-Smirnov test statistic is

$$\begin{aligned}
D_n &= \sqrt{n}\sup_y |F_n(y) - F_0(y)| = \max(D_n^+, D_n^-), \text{ where} \\
D_n^+ &= \max_{1\le i\le n}\left[\frac{i}{n} - F_0(Y_{(i)})\right] \text{ and } D_n^- = \max_{2\le i\le n}\left[F_0(Y_{(i)}) - \frac{i-1}{n}\right]
\end{aligned}$$

and $F_n(y)$ is the empirical distribution function. It can be shown that when H_0 is true, then the distribution of D_n does not depend on F_0. Tables of percentiles of D_n are available in many textbooks.

2.4.1.5 *Cramer-Von Mises*

Again, consider a test of $\mathrm{H}_0\colon F_Y(y) = F_0(y)$ against the alternative $\mathrm{H}_a\colon F_Y(y) \neq F_0(y)$. Denote the empirical cdf of the standardized sample values $Z_i = (Y_i - \overline{Y})/S$ by $\widehat{F}_Z(z)$. The Cramer-Von Mises test statistic is

$$C_n^2 = n\int_{-\infty}^{\infty}\left[\widehat{F}_Z(z) - \Phi(z)\right]^2\phi(z)\,dz = \sum_{i=1}^{n}\left[\Phi(Z_{(i)}) - \left(\frac{2i-1}{2n}\right)\right]^2 + \frac{1}{12n},$$

where $\Phi(z)$ and $\phi(z)$ are the cdf and the pdf of the standard normal distribution. The hypothesis of normality is rejected for large values of C_n^2.

2.4.1.6 *Anderson-Darling*

For a third time, consider a test of $\mathrm{H}_0\colon F_Y(y) = F_0(y)$ against the alternative $\mathrm{H}_a\colon F_Y(y) \neq F_0(y)$. Denote the empirical cdf of the standardized sample values $Z_i = (Y_i - \overline{Y})/S$ by $\widehat{F}_Z(z)$. The Anderson-Darling test statistic is

$$\begin{aligned} A_n^2 &= n\int_{-\infty}^{\infty} \frac{\left[\widehat{F}_Z(z) - \Phi(z)\right]^2}{\Phi(z)[1-\Phi(z)]}\phi(z)\,dz \\ &= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left\{\ln\left[\Phi(Z_{(i)})\right] + \ln\left[1-\Phi(Z_{(n-i+1)})\right]\right\}. \end{aligned}$$

The hypothesis of normality is rejected for large values of A_n^2. The upper 0.05 and 0.01 critical values are, approximately,

$$A_{n,0.05}^2 = 0.7514\left(1 - \frac{0.795}{n} - \frac{0.89}{n^2}\right) \text{ and } A_{n,0.01}^2 = 1.0348\left(1 - \frac{1.013}{n} - \frac{0.93}{n^2}\right).$$

2.4.2 Multivariate Procedures

Assume that $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ is a random sample from a d dimensional population with pdf $f_{\mathbf{y}}(\mathbf{y})$ and cdf $F_{\mathbf{y}}(\mathbf{y})$. Denote the sample mean and variance by $\overline{\mathbf{y}}$ and $\mathbf{S}$, respectively.

2.4.2.1 *QQ Plot of Squared Mahalanobis Distance*

The squared Mahalanobis distance from $\mathbf{y}_i$ to $\overline{\mathbf{y}}$ is

$$D_i^2 = (\mathbf{y}_i - \overline{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}}).$$

It can be shown that if the data have been sampled from a multivariate normal population, then

$$\frac{n}{(n-1)^2}D_i^2 \sim \text{Beta}\left(\frac{d}{2}, \frac{n-d-1}{2}\right).$$

One graphical test of multivariate is a QQ plot of the ordered D_i^2 statistics against the quantiles of the beta distribution with parameters $d/2$ and $(n-d-1)/2$. This plot can be constructed in SAS. I will give instructions in class.

2.4.2.2 *Multivariate Outliers*

One simple test to determine whether there are any outliers is to examine the maximum D_i^2 value. Let $D_{\max}^2 = \max D_i^2$. Using the Bonferroni inequality, the hypothesis that no outliers exist can be rejected if

$$F_{\max} = \left(\frac{n-d-1}{d}\right)\left[\frac{1}{1 - nD_{\max}^2/(n-1)^2} - 1\right] \geq F_{d,n-d-1}^{1-\alpha/n}.$$

This test suffers from masking and swamping. An outlier is masked if it can not be detected unless certain other observations are deleted from the data set.

2.4.2.3 *Mardia's Coefficients of Skewness and Kurtosis*

Mardia has generalized skewness and kurtosis coefficients to multivariate distributions. The coefficients and their estimators are described on pages 96–99 in Rencher (2002).

2.4.2.4 *Henze-Zirkler Invariant Test*

Consider a test of $\mathrm{H}_0\colon \mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ against the alternative $\mathrm{H}_a\colon \mathbf{y} \not\sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Denote the empirical characteristic function of the standardized sample values $\mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{y}_i - \overline{\mathbf{y}})$ by $\psi_n(t)$. That is,

$$\psi_n(t) = \frac{1}{n}\sum_{j=1}^{n}\exp\{i\mathbf{t}'\mathbf{z}_i\},$$

where $i^2 = -1$. Denote the characteristic function of $\mathrm{N}(\mathbf{0}, \mathbf{I}_d)$ as $\psi_0(t)$. That is,

$$\psi_0(t) = \mathrm{E}\left(\exp\{i\mathbf{t}'\mathbf{u}\}\right) = \exp\{-\mathbf{t}'\mathbf{t}/2\},$$

where $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_d)$. The Henze-Zirkler test statistic is

$$D_{n,\beta} = \int_{\mathbf{R}^d} \left|\psi_n(\mathbf{t}) - \psi_0(\mathbf{t})\right|^2 \phi_\beta(\mathbf{t})\, d\mathbf{t},$$

where $\phi_\beta(\mathbf{t})$ is the pdf of $\mathrm{N}(\mathbf{0}, \beta_n^2 \mathbf{I}_d)$. Henze and Zirkler (*Communications in Statistics — Theory and Methods*, 1990, **19**, 3595–3617) give an equation for β and describe how the test statistic can be computed. The distribution of the test statistic can be approximated by a log normal distribution. The null hypothesis is rejected for small values of the test statistic. This test can be performed in SAS.

2.4.3 Testing Normality in SAS

1. Histogram and Kernel Smoothed Density Plot

```
proc univariate data = dataset plots noprint;
    var Y1;
    title 'Smoothed Histogram of Y1';
    histogram Y1 /kernel(l=1 color = black);
    inset mean std skewness kurtosis;
run;
```

2. Normal Probability Plot, Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling

```
proc univariate data = dataset normal;
 var Y1-Y4;
 qqplot Y1-Y4 /normal(mu=est sigma=est) cframe=ligr
                 pctlaxis(grid lgrid=35 label='Normal Percentiles');
    inset mean std / cfill=white format=3.0 header='Normal Parameters'
                     position=(95,10) refpoint=br;
run;
```

3. QQ Plot of Scaled D_i^2 fit to Beta, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling

```
proc iml;
 small=1.e-10;

/*----------------------------------------*/
/* Module to compute the rank of a matrix  */
/*----------------------------------------*/
start rankM(A) global(small);
  m = nrow(A); n = ncol(A);
  call svd(U,S,V,A);
  if m = 1
    then S = S[1];
    else if m = 0 then S = {0};
  tol = max(S) * small;
  r = max(loc(S>tol)); /*rank of A = number of nonzero singular values */
  return(r);
finish rankM;

 use new;
 read all var {y1 y2 y3} into Y;
 read all var {x1 x2 x3} into X1;
 reset noprint;
```

```
 n=nrow(Y); d=ncol(Y);
 X=J(n,1) || X1;
 r=rankM(X);
 H=X*ginv(X‘*X)*X‘;
 E=Y-H*Y;
 S=(E‘*E)/(n-r);
 A=E*inv(S)*E‘;
 DD=vecdiag(A);
 h1=vecdiag(I(n)-H);
 u=(DD#H1##(-1))/(n-r);
*;
*   Under multivariate normality, the entries of u are each;
*   distributed as a beta random variable with parameters;
*   d/2 and (n-r-d)/2, where r = rank(X);
*   For this data set d/2=1.5 and (n-r-d)/2 = 24;
*;

 Dn=max(DD);
 Un=max(u);
 F_max=((n-r-d)/d)*Un/(1-Un);
 prob_F =n*(1- probf(F_max,d,n-r-d));
 print Dn Un F_max prob_F;
 create tdata from u [colname = 'u'];
 append from u;
 close tdata;


data total;
 merge res_out tdata;

proc univariate data = total;
 var u;
*;
*  If d = 3, n = 50, and r=3 then alpha = 1.5 and beta = 22;
*;
 qqplot u/beta(alpha=1.5 beta=22 threshold=0 scale=1);
 histogram u / beta(alpha=1.5 beta=22);
   inset mean (5.3) std = 'Std Dev' (5.3) Skewness (5.3)
   Kurtosis (5.3) /header = 'Summary Statistics' pos = nw;
   title1 'Plot of Mahalanobis Distances';
run;
```

4. Henze-Zirkler Invariant Test and Mardia's Multivariate Skewness and Kurtosis

```
proc model;
    parms b01 b11 b21 b31 b02 b12 b22 b32 b03 b13 b23 b33;
    instrument x1 x2 x3;
    Y1 = b01 +b11*x1 + b21*x2 + b31*x3;
    Y2 = b02 +b12*x1 + b22*x2 + b32*x3;
    Y3 = b03 +b13*x1 + b23*x2 + b33*x3;
    fit Y1-Y3 /normal;
run;
```

2.5 TRANSFORMATIONS TO NORMALITY

Consider the linear model $\text{vec}(\mathbf{Y}) \sim \text{N}\left[\text{vec}(\mathbf{XB}), \mathbf{\Sigma} \otimes \mathbf{I}_n\right]$. This model requires that the rows of $\mathbf{Y}$, say $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ be independently distributed as normal random vectors with equal variance and that $\text{E}(\mathbf{Y}) \in \mathcal{R}(\mathbf{X})$.

It often happens that data do not conform to these assumptions. A typical problem is that the covariance matrix of the $\mathbf{y}_i$s is not constant. Sometimes $\text{Var}(\mathbf{y}_i)$ is a function of $\text{E}(\mathbf{y}_i)$. If this is the case, then the investigator could use weighted regression with weights $\mathbf{W}_i \propto [\text{Var}(\mathbf{y}_i)]^{-1}$ or could try to transform the data to better meet the assumptions. A method for selecting a transformation are described below.

2.5.1 The Box-Cox Family of Transformations: Univariate Approach

One approach to transforming the data is to find a set of d transformations so that if the j^{th} transformation is applied to the j^{th} column of $\mathbf{Y}$ then the result will be a column vector that satisfies the usual assumptions. In particular, the Box-Cox family of transformations (Box and Cox, 1964) could be applied separately to each column of $\mathbf{Y}$.

Let $\mathbf{u}_j$ be the j^{th} column of $\mathbf{Y}$. Then, the transformed variables are

$$z_{ij} = \frac{u_{ij}^{\lambda_j} - 1}{\lambda_j}.$$

The goal is to choose λ_j so that $\mathbf{z}_j \sim \text{N}(\mathbf{X}\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I})$. It can be shown by using L'Hopital's rule, that

$$\lim_{\lambda_j \to 0} z_{ij} = \ln(u_{ij}).$$

To choose λ_j, assume that the random vector $\mathbf{z}_j$ has distribution $\mathbf{z}_j \sim \text{N}_n(\mathbf{X}\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I})$. Make the transformation from $\mathbf{z}_j$ to $\mathbf{u}_j$. The Jacobian of the transformation is $\prod_{i=1}^n u_{ij}^{\lambda_j - 1}$. Accordingly, the random vector $\mathbf{u}_j$ has density function

$$f(\mathbf{u}_j) = \frac{\exp\left\{-\frac{1}{2\sigma_j^2} q(\mathbf{u}_j)\right\}}{(2\pi\sigma_j^2)^{\frac{n}{2}}} \left(\prod_{i=1}^n u_{ij}^{\lambda_j - 1}\right),$$

where

$$q(\mathbf{u}_j) = \sum_{i=1}^n \left(\frac{u_{ij}^{\lambda_j} - 1}{\lambda_j} - \mathbf{x}_i'\boldsymbol{\beta}_j\right)^2 = (\mathbf{z}_j - \mathbf{X}\boldsymbol{\beta}_j)'(\mathbf{z}_j - \mathbf{X}\boldsymbol{\beta}_j).$$

The parameter λ_j can be estimated by maximizing the likelihood function. In practice, it is easier to maximize the log likelihood. For fixed λ_j, the MLE's of σ_j^2 and $\boldsymbol{\beta}_j$ are known to be

$$\widehat{\sigma}_{\lambda_j}^2 = \frac{\mathbf{z}_j'(\mathbf{I} - \mathbf{H})\mathbf{z}_j}{n},$$

and

$$\widehat{\boldsymbol{\beta}}_{\lambda_j} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{z}_j,$$

where $\mathbf{H} = \text{ppo}(\mathbf{X})$. Hence, omitting the constants, the profile log likelihood function (i.e., the likelihood function maximized over σ_j^2 and $\boldsymbol{\beta}_j$) for fixed λ_j is

$$L(\lambda_j) = -\frac{n}{2}\ln(\widehat{\sigma}_{\lambda_j}^2) + n(\lambda_j - 1)\overline{w},$$

where

$$\overline{w} = n^{-1}\sum_{i=1}^n \ln(u_{ij}).$$

Note that $e^{\overline{w}}$ is the geometric mean of response vector. Alternatively, let

$$t_i = \frac{z_{ij}}{g^{\lambda_j - 1}},$$

where

$$g = \left(\prod_{i=1}^n u_{ij}\right)^{\frac{1}{n}} = e^{\overline{w}}$$

is the geometric mean. Then, $L(\lambda_j)$ simplifies to

$$L(\lambda_j) = -\frac{n}{2}\ln(s_{\lambda_j}^2),$$

where

$$s^2_{\lambda_j} = \frac{\mathbf{t}'(\mathbf{I}-\mathbf{H})\mathbf{t}}{n} \text{ and } \mathbf{t} = \mathbf{z}_j \times \frac{1}{g^{\lambda_j - 1}}.$$

To maximize the likelihood function, one need only find the value of λ_j which minimizes $s^2_{\lambda_j}$. Large sample confidence intervals for λ_j, in the Box-Cox family can be constructed by inverting the generalized likelihood ratio test. A value, λ_0, is inside the $(1-\alpha)100\%$ confidence interval if

$$-2\ln\left[\frac{e^{L(\lambda_0)}}{e^{L(\hat{\lambda}_j)}}\right] \leq \chi^2_{1-\alpha,1}.$$

Thus, a $100(1-\alpha)\%$ confidence interval for λ_j consists of all values, λ_0 that satisfy

$$n\ln\left[\frac{SSE(\lambda_0)}{SSE(\widehat{\lambda}_j)}\right] \leq \chi^2_{1-\alpha,1},$$

where $SSE(\lambda_0)$ is the error sum of squares,

$$SSE(\lambda_0) = \mathbf{t}'(\mathbf{I}-\mathbf{H})\mathbf{t},$$

computed using $\lambda_j = \lambda_0$, and $SSE(\widehat{\lambda}_j)$ is the error sum of squares computed using $\lambda_j = \widehat{\lambda}_j$, the MLE of λ_j.

2.5.2 The Box-Cox Family of Transformations: Multivariate Approach

In the approach in the previous section, the columns of $\mathbf{Y}$ were transformed one at a time. It is well known that marginal normality does not imply joint normality so a better approach might be to choose the transformation powers simultaneously.

Let $\mathbf{Z}$ be the matrix of transformed responses, where

$$z_{ij} = \frac{y_{ij}^{\lambda_j} - 1}{\lambda_j}.$$

If $\text{vec}(\mathbf{Z}) \sim \text{N}\left[\text{vec}(\mathbf{XB}, \mathbf{\Sigma} \otimes \mathbf{I}_n\right]$, then the pdf of $\mathbf{Y}$ is

$$f(\mathbf{Y}; \mathbf{\Sigma}, \mathbf{B}, \boldsymbol{\lambda}) = \frac{\exp\left\{-\frac{1}{2}\,\text{tr}\left[(\mathbf{Z}-\mathbf{XB})'(\mathbf{Z}-\mathbf{XB})\mathbf{\Sigma}^{-1}\right]\right\}}{(2\pi)^{nd/2}|\mathbf{\Sigma}|^{n/2}} \prod_{i=1}^{n}\prod_{j=1}^{d} y_{ij}^{\lambda_j - 1}.$$

The parameters $\mathbf{B}$, $\mathbf{\Sigma}$, and $\boldsymbol{\lambda} = \begin{pmatrix}\lambda_1 & \dots & \lambda_d\end{pmatrix}'$ can be estimated by maximizing the likelihood function.

A variant of this approach is to require that all λ values be identical; i.e., $\boldsymbol{\lambda} = \mathbf{1}_d\lambda$ for some scalar λ. This constraint is sensible in repeated measures or longitudinal studies where the d measures represent the same variable observed on d occasions.

2.6 CORRELATION AND REGRESSION

2.6.1 Correlation: Population Parameters

Consider the random d-vector $\mathbf{y} \sim (\boldsymbol{\mu}, \mathbf{\Sigma})$. Denote the jk^{th} element of $\mathbf{\Sigma}$ by σ_{jk}. Be careful, $\text{var}(y_j) = \sigma_j^2 = \sigma_{jj}$, not σ_{jj}^2. The correlation between y_j and y_k is defined as

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}.$$

By the Cauchy-Schwartz inequality, $\rho_{jk}^2 \leq 1$. The quantity ρ_{jk}^2 is called the coefficient of determination between variables j and k.

Define $\mathbf{D}$ by $\mathbf{D} = \text{diag}(\mathbf{\Sigma})$. Then, the matrix of correlations is

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{D}^{-\frac{1}{2}}.$$

It can be shown that $0 \leq |\mathbf{R}| \leq 1$.

2.6.2 Correlation: Sample Statistics

Let $\mathbf{Y}$ be an $n \times d$ matrix for which the rows of $\mathbf{Y}$ are a random sample from $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Denote the j^{th} column of $\mathbf{Y}$ by $\mathbf{y}_j$, denote the ij^{th} entry of $\mathbf{Y}$ by y_{ij} and denote the mean of the j^{th} column by $\bar{y}_{\cdot j}$. The usual sample estimate of ρ_{jk} is

$$\begin{aligned} r_{jk} &= \frac{\sum_{i=1}^{n}(y_{ij}-\bar{y}_{\cdot j})(y_{ik}-\bar{y}_{\cdot k})}{\sqrt{\sum_{i=1}^{n}(y_{ij}-\bar{y}_{\cdot j})^2 \sum_{i=1}^{n}(y_{ik}-\bar{y}_{\cdot k})^2}} \\ &= \frac{(\mathbf{y}_j - \mathbf{1}_n\bar{y}_{\cdot j})'(\mathbf{y}_k - \mathbf{1}_n\bar{y}_{\cdot k})}{\sqrt{[(\mathbf{y}_j - \mathbf{1}_n\bar{y}_{\cdot j})'(\mathbf{y}_j - \mathbf{1}_n\bar{y}_{\cdot j})]\,[(\mathbf{y}_k - \mathbf{1}_n\bar{y}_{\cdot k})'(\mathbf{y}_k - \mathbf{1}_n\bar{y}_{\cdot k})]}} \\ &= \frac{\mathbf{y}_j'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_k}{\sqrt{\left[\mathbf{y}_j'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_j\right]\left[\mathbf{y}_k'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_k\right]}} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}, \end{aligned}$$

where $\mathbf{H}_1 = \text{ppo}(\mathbf{1}_n)$ and s_{jk} is the jk^{th} entry in

$$\mathbf{S} = \frac{\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y}}{(n-1)}.$$

Let

$$\mathbf{u}_j = \mathbf{y}_j - \mathbf{1}_n\bar{y}_{\cdot j} = (\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_j$$

for $j = 1, \ldots, d$. Let θ_{jk} be the angle between $\mathbf{u}_j$ and $\mathbf{u}_k$. Then

$$\cos(\theta_{jk}) = \frac{\mathbf{u}_j'\mathbf{u}_k}{\sqrt{\left[\mathbf{u}_j'\mathbf{u}_j\right]\left[\mathbf{u}_k'\mathbf{u}_k\right]}} = r_{jk}.$$

The matrix of sample correlations, $\widehat{\mathbf{R}} = \{r_{jk}\}$, can be computed as follows:

$$\widehat{\mathbf{R}} = \widehat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{S}\widehat{\mathbf{D}}^{-\frac{1}{2}},$$

where

$$\widehat{\mathbf{D}} = \text{diag}(\mathbf{S}), \text{ and } \mathbf{H}_1 = \text{ppo}(\mathbf{1}_n).$$

Inferences about ρ_{ij} can be made using the following distributional results.

Theorem 2.4 *If $\mathbf{Y}$ is normally distributed and $\rho_{ij} = 0$, then*

$$t = \frac{r_{ij}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}} \sim t_{n-2}.$$

The null hypothesis $\text{H}_0\colon \rho_{ij} = 0$ *can be rejected in favor of the alternative* $\text{H}_a\colon \rho_{ij} \neq 0$ *if* $|t| \geq t_{n-2}^{1-\alpha/2}$.

Proof: The conditional distribution of $\mathbf{y}_i$ given $\mathbf{y}_j = \ddot{\mathbf{y}}_j$ is

$$\mathbf{y}_i|\mathbf{y}_j \sim \text{N}(\mathbf{1}_n\beta_0 + \ddot{\mathbf{y}}_j\beta_1, \sigma_{ii\cdot j}\mathbf{I}_n),$$

where $\beta_0 = \mu_i - \beta_1\mu_j$; $\beta_1 = \sigma_{ij}/\sigma_{jj}$; and $\sigma_{ii\cdot j} = \sigma_{ii} - \sigma_{ij}\sigma_{jj}^{-1}\sigma_{ji} = \sigma_{ii}(1-\rho_{ij}^2)$. The usual t statistic for testing $\text{H}_0\colon \beta_1 = 0$ *is*

$$t = \frac{\widehat{\beta}_1}{\widehat{SE}(\widehat{\beta}_1)} \sim t_{n-2,\lambda}, \text{ where } \lambda = \frac{\beta_1^2(n-1)s_{jj}}{2\sigma_{ii\cdot j}^2} = \frac{(n-1)s_{jj}\rho_{ij}^2}{2\sigma_{jj}(1-\rho_{ij}^2)}.$$

Using the annihilator, it is readily shown that

$$\widehat{\beta}_1 = \left[\ddot{\mathbf{y}}_j'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{y}}_j\right]^{-1}\ddot{\mathbf{y}}_j'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_i = s_{jj}^{-1}s_{ji} = r_{ij}\sqrt{\frac{s_{ii}}{s_{jj}}}, \text{ where } \mathbf{H}_1 = \text{ppo}(\mathbf{1}_n);$$

$$\text{Var}(\widehat{\beta}_1) = \sigma_{ii\cdot j}\left[\ddot{\mathbf{y}}_j'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{y}}_j\right]^{-1} = \frac{\sigma_{ii\cdot j}}{(n-1)s_{jj}} \text{ and}$$

$$\widehat{\sigma}_{ii\cdot j} = \frac{\mathbf{y}_i'(\mathbf{I}_n - \mathbf{H})\mathbf{y}_i}{n-2} = \frac{\mathbf{y}_i'(\mathbf{I}_n - \mathbf{H}_1 - \mathbf{H}_{2\cdot 1})\mathbf{y}_i}{n-2} = \left(\frac{n-1}{n-2}\right) s_{ii}(1-r_{ij}^2).$$

Accordingly,

$$t = \frac{r_{ij}\sqrt{\dfrac{s_{ii}}{s_{jj}}}}{\sqrt{\dfrac{s_{ii}(1-r_{ij}^2)}{(n-2)s_{jj}}}} = \frac{r_{ij}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}}.$$

If $\rho_{ij} = 0$, then $\beta_1 = 0$ and t has a central t distribution with $n-2$ degrees of freedom.

□

Theorem 2.5 (Fisher's Z) *If $\mathbf{Y}$ is normally distributed and n is not small, then*

$$\frac{1}{2}\ln\left(\frac{1+r_{ij}}{1-r_{ij}}\right) \dot{\sim} \mathrm{N}\left[\xi_{ij}, (n-3)^{-1}\right],$$

where

$$\xi_{ij} = \frac{1}{2}\ln\left(\frac{1+\rho_{ij}}{1-\rho_{ij}}\right).$$

The statistic

$$Z_{ij} = \frac{1}{2}\ln\left(\frac{1+r_{ij}}{1-r_{ij}}\right)$$

is called Fisher's Z. The endpoints of an approximate $100(1-\alpha)\%$ confidence interval for ξ_{ij} are given by

$$Z_{ij} \pm \frac{z^*_{1-\alpha/2}}{\sqrt{n-3}},$$

where $z^*_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. Back transforming from ξ_{ij} to ρ_{ij} yields an approximate $100(1-\alpha)\%$ confidence interval for ρ_{ij}:

$$\frac{\exp\left\{2\left(Z_{ij} - z^*_{1-\alpha/2}/\sqrt{n-3}\right)\right\} - 1}{\exp\left\{2\left(Z_{ij} - z^*_{1-\alpha/2}/\sqrt{n-3}\right)\right\} + 1} \le \rho_{ij} \le \frac{\exp\left\{2\left(Z_{ij} + z^*_{1-\alpha/2}/\sqrt{n-3}\right)\right\} - 1}{\exp\left\{2\left(Z_{ij} + z^*_{1-\alpha/2}/\sqrt{n-3}\right)\right\} + 1}.$$

□

2.6.3 Multiple Correlation: Population Parameter

Consider the random d-vector $\mathbf{y} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} > 0$. Conformably partition $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \mathbf{y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{21}' \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where y_1, μ_1, and $\sigma_{11} = \sigma_1^2$ are scalars. The squared multiple correlation between y_1 and $\mathbf{y}_2$ is defined as the maximum squared correlation between y_1 and a linear combination of $\mathbf{y}_2$. That is

$$\rho_{12}^2 = \max_{\mathbf{t}} \left[\mathrm{corr}(y_1, \mathbf{t}'\mathbf{y}_2)\right]^2.$$

Theorem 2.6 *The squared multiple correlation between y_1 and $\mathbf{y}_2$ is*

$$\rho_{12}^2 = \left[\mathrm{corr}(y_1, \mathbf{t}'\mathbf{y}_2)\right]^2 = \frac{\boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}}{\sigma_{11}},$$

where $\mathbf{t} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$.

□

2.6.4 Multiple Correlation: Sample Statistics

Let $\mathbf{Y}$ be an $n \times d$ matrix for which the rows of $\mathbf{Y}$ are a random sample from $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let

$$\mathbf{S} = \frac{\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y}}{(n-1)} = \begin{pmatrix} s_{11} & \mathbf{s}_{21}' \\ \mathbf{s}_{21} & \mathbf{S}_{22} \end{pmatrix} \text{ where } \mathbf{H}_1 = \text{ppo}(\mathbf{1}_n).$$

The sample estimator of ρ_{12}^2 is

$$R_{12}^2 = \frac{\mathbf{s}_{21}'\mathbf{S}_{22}^{-1}\mathbf{s}_{21}}{s_{11}}.$$

Theorem 2.7 *If $\mathbf{Y}$ has a multivariate normal distribution and $\rho_{12}^2 = 0$, then*

$$F = \left(\frac{n-d}{d-1}\right)\left(\frac{R_{12}^2}{1-R_{12}^2}\right) \sim F_{d-1,n-d}.$$

The null hypothesis $\mathrm{H}_0\colon \rho_{12} = 0$ can be rejected in favor of $\mathrm{H}_a\colon \rho_{12} \neq 0$ if $F \geq F_{d-1,n-d}^{1-\alpha}$.

Proof: Partition $\mathbf{Y}$ as $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{Y}_2 \end{pmatrix}$, where $\mathbf{Y}_2$ is $n \times (d-1)$. The conditional distribution of $\mathbf{y}_i$ given $\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2$ is

$$\mathbf{y}_1 | \mathbf{Y}_2 \sim \mathrm{N}(\mathbf{1}_n\beta_0 + \ddot{\mathbf{Y}}_2\boldsymbol{\beta}_1, \sigma_{11\cdot 2}\mathbf{I}_n),$$

or, equivalently,

$$\mathbf{y}_1 | \mathbf{Y}_2 \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),$$

where

$$\begin{aligned}
\beta_0 &= \mu_1 - \boldsymbol{\beta}_1'\boldsymbol{\mu}_2; \quad \boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}; \\
\sigma_{11\cdot 2} &= \sigma_{11} - \boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} = \sigma_{11}(1-\rho_{12}^2); \\
\boldsymbol{\beta} &= \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \ddot{\mathbf{Y}}_2 \end{pmatrix}; \text{ and} \\
\sigma^2 &= \sigma_{11\cdot 2}.
\end{aligned}$$

The model comparison likelihood ratio test for testing $\mathrm{H}_0\colon \boldsymbol{\beta}_1 = \mathbf{0}$ is to reject H_0 if $F \geq F_{d-1,n-d}^{1-\alpha}$, where

$$\begin{aligned}
F &= \left(\frac{n-d}{d-1}\right)\frac{\mathbf{y}_1'(\mathbf{H} - \mathbf{H}_1)\mathbf{y}_1}{\mathbf{y}_1'(\mathbf{I}_n - \mathbf{H})\mathbf{y}_1} \text{ where } \mathbf{H} = \text{ppo}(\mathbf{X}) \\
&= \left(\frac{n-d}{d-1}\right)\frac{\mathbf{y}_1'\mathbf{H}_{2\cdot 1}\mathbf{y}_1}{\mathbf{y}_1'(\mathbf{I}_n - \mathbf{H}_1 - \mathbf{H}_{2\cdot 1})\mathbf{y}_1} \\
&= \left(\frac{n-d}{d-1}\right)\frac{R_{12}^2}{(1-R_{12}^2)}.
\end{aligned}$$

Furthermore, conditional on $\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2$, the distribution of F is $F \sim F_{d-1,n-d,\lambda}$, where

$$\lambda = \frac{\boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2'(\mathbf{I} - \mathbf{H}_1)\mathbf{Y}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}}{2\sigma_{11}(1-\rho_{12}^2)} = (n-1)\frac{\boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\mathbf{S}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}}{2\sigma_{11}(1-\rho_{12}^2)}.$$

Note that

$$\mathrm{E}(\lambda) = (n-1)\frac{\rho_{12}^2}{2(1-\rho_{12}^2)}.$$

Also, under $\mathrm{H}_0\colon \rho_{12} = 0$, the noncentrality parameter goes to zero and the test statistic has a central F distribution.

□

2.6.5 More on Conditional Distributions

In this section additional details on the conditional distribution of $\mathbf{y}_1$ given $\ddot{\mathbf{Y}}_2$ are described. The results in this section provide an alternative proof of Theorem 2.7.

Let $\mathbf{Y}$ be an $n \times d$ matrix for which the rows of $\mathbf{Y}$ are a random sample from $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. That is,

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\mu}' + \mathbf{E}, \text{ where } \mathrm{vec}(\mathbf{E}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n).$$

Equivalently,

$$\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}\left[(\mathbf{I}_d \otimes \mathbf{1}_n)\boldsymbol{\mu}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n\right].$$

Partition $\mathbf{Y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{Y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{21}' \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{y}_1$ is $n \times 1$, and μ_1 and σ_{11} are scalars. From previous results, it is known that

$$\mathbf{y}_1 | (\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2) \sim \mathrm{N}\left[\mathbf{1}_n \mu_1 + (\ddot{\mathbf{Y}}_2 - \mathbf{1}_n \boldsymbol{\mu}_2')\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}, \sigma_{11\cdot 2}\mathbf{I}_n\right],$$

where

$$\sigma_{11\cdot 2} = \sigma_{11} - \boldsymbol{\sigma}_{21}'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}.$$

Rearranging terms yields the regression model

$$\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \ddot{\mathbf{Y}}_2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \mu_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \\ \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \end{pmatrix}, \text{ and } \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_{11\cdot 2}\mathbf{I}_n).$$

Using annihilator results, the MLE (OLS) of $\boldsymbol{\beta}_2$ is

$$\hat{\boldsymbol{\beta}}_2 = \left[\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\right]^{-1} \ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_1 = \mathbf{S}_{22}^{-1}\mathbf{s}_{21},$$

where $\mathbf{H}_1 = \mathrm{ppo}(\mathbf{1}_n)$. The conditional distribution of $\hat{\boldsymbol{\beta}}_2$ given $\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2$ is

$$\hat{\boldsymbol{\beta}}_2 | (\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2) \sim \mathrm{N}\left\{\boldsymbol{\beta}_2, \sigma_{11\cdot 2}\left[\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\right]^{-1}\right\}.$$

That is,

$$\hat{\boldsymbol{\beta}}_2 | (\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2) \sim \mathrm{N}\left[\boldsymbol{\beta}_2, \sigma^2\mathbf{S}_{22}^{-1}/(n-1)\right],$$

where $\sigma^2 = \sigma_{11\cdot 2}$. The usual estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\mathbf{y}_1'(\mathbf{I}_n - \mathbf{H})\mathbf{y}_1}{n-d},$$

where $\mathbf{H} = \mathrm{ppo}(\mathbf{X})$ and $\mathrm{rank}(\mathbf{X}) = 1 + (d-1) = d$. Write $\mathbf{X}$ as $\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \ddot{\mathbf{Y}}_2 \end{pmatrix}$. Recall that $\mathbf{H}$ can be decomposed as

$$\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_{2\cdot 1},$$

where

$$\begin{aligned} \mathbf{H}_1 &= \mathrm{ppo}(\mathbf{1}_n) = n^{-1}\mathbf{J}_n^n \text{ and} \\ \mathbf{H}_{2\cdot 1} &= \mathrm{ppo}\left\{[\mathbf{I}_n - \mathrm{ppo}(\mathbf{1}_n)]\,\ddot{\mathbf{Y}}_2\right\} = (\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\left[\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\right]^{-1}\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1) \end{aligned}$$

Accordingly, $\hat{\sigma}^2$ can be written as

$$\hat{\sigma}^2 = \frac{\mathbf{y}_1'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_1 - \mathbf{y}_1'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\left[\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\right]^{-1}\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\mathbf{y}_1}{n-d}$$

$$= (n-1)\frac{s_{11} - \mathbf{s}_{21}'\mathbf{S}_{22}^{-1}\mathbf{s}_{21}}{n-d} = \frac{(n-1)s_{11}}{n-d}\left(1 - R_{12}^2\right).$$

Conditional on $\mathbf{Y}_2 = \ddot{\mathbf{Y}}_2$, the likelihood ratio test of $\mathrm{H}_0\colon \boldsymbol{\beta}_2 = \mathbf{0}$ versus $\mathrm{H}_a\colon \boldsymbol{\beta}_2 \neq \mathbf{0}$ is to reject H_0 if $F \geq F_{d-1,n-d}^{1-\alpha}$, where

$$F = \frac{\hat{\boldsymbol{\beta}}_2'\ddot{\mathbf{Y}}_2'(\mathbf{I}_n - \mathbf{H}_1)\ddot{\mathbf{Y}}_2\hat{\boldsymbol{\beta}}_2}{(d-1)\hat{\sigma}^2} = \left(\frac{n-d}{d-1}\right)\left(\frac{R_{12}^2}{1-R_{12}}\right).$$

Note that the unconditional distribution of F under H_0 is $F_{d-1,n-d}$. Also note that

$$R_{12}^2 = \frac{\mathbf{s}_{21}'\mathbf{S}_{22}^{-1}\mathbf{s}_{21}}{s_{11}} = \frac{R(\beta_1, \boldsymbol{\beta}_2) - R(\beta_1)}{\mathbf{y}_1'\mathbf{y}_1 - R(\beta_1)},$$

which is the usual R^2 in regression models.

2.6.6 Partial Correlation

Consider two random vectors, $\mathbf{y}\colon p \times 1$ and $\mathbf{x}\colon q \times 1$ with joint distribution

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx}\\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx}\end{pmatrix}\right].$$

The covariance matrix for $\mathbf{y}|(\mathbf{x} = \ddot{\mathbf{x}})$ is called the partial covariance matrix and is given by

$$\mathrm{Var}[\mathbf{y}|(\mathbf{x} = \ddot{\mathbf{x}})] = \boldsymbol{\Sigma}_{yy\cdot x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}.$$

Let $\mathbf{D}_{yy\cdot x} = \mathrm{diag}(\boldsymbol{\Sigma}_{yy\cdot x})$. Then, the partial correlation matrix for $\mathbf{y}$ given $\mathbf{x}$ is

$$\mathbf{R}_{yy\cdot x} = \mathbf{D}_{yy\cdot x}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{yy\cdot x}\mathbf{D}_{yy\cdot x}^{-\frac{1}{2}}.$$

Note that the partial covariances and partial correlations do not depend on $\ddot{\mathbf{x}}$. If $p = 2$ and $q = 1$, then

$$\rho_{y_1,y_2|x} = \frac{\rho_{y_1,y_2} - \rho_{y_1,x}\rho_{y_2,x}}{\sqrt{(1-\rho_{y_1,x}^2)(1-\rho_{y_2,x}^2)}}.$$

2.6.7 Prediction & Regression: Population Parameters

2.6.7.1 *Best Predictor (BP)*

Consider two random vectors, $\mathbf{x}$ and $\mathbf{y}$, with joint density $f(\mathbf{x}, \mathbf{y})$. The density need not be normal. Suppose that $\mathbf{x}$ can be observed, but $\mathbf{y}$ can not be observed. We wish to predict $\mathbf{y}$ based on the observed $\mathbf{x}$. Denote the predicted value of $\mathbf{y}$ by $\hat{\mathbf{g}} = \hat{\mathbf{g}}(\mathbf{x})$ (random variable) or $\hat{\mathbf{g}} = \hat{\mathbf{g}}(\ddot{\mathbf{x}})$ (realization). The best predictor is defined as the function $\hat{\mathbf{g}}$ which minimizes

$$MSE(\hat{\mathbf{g}}) = \mathrm{E}\left[(\mathbf{y} - \hat{\mathbf{g}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \hat{\mathbf{g}})\right],$$

where $\boldsymbol{\Omega}$ is a positive definite matrix (e.g., a covariance matrix).

Theorem 2.8 *The best predictor is* $\hat{\mathbf{g}} = \mathrm{E}(\mathbf{y}|\mathbf{x})$.

□

Corollary. The best predictor is unbiased: $\mathrm{E}[\mathrm{E}(\mathbf{y}|\mathbf{x})] = \mathrm{E}(\mathbf{y})$.

2.6.7.2 *Regression Under Normality*

Consider the random $(p+1)$-vector $\mathbf{z} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} > 0$. Conformably partition $\mathbf{z}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\mathbf{z} = \begin{pmatrix}y\\ \mathbf{x}\end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix}\mu_y\\ \boldsymbol{\mu}_x\end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_y^2 & \boldsymbol{\Sigma}_{yx}\\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx}\end{pmatrix},$$

where y, μ_y, and σ_y^2 are scalars. Suppose that we wish to predict y after observing $\mathbf{x} = \ddot{\mathbf{x}}$. The best predictor is

$$\mathrm{E}(y|\mathbf{x} = \ddot{\mathbf{x}}) = \mu_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\ddot{\mathbf{x}} - \boldsymbol{\mu}_x) = \ \beta_0 + \boldsymbol{\beta}_1'\ddot{\mathbf{x}},$$

where

$$\beta_0 = \mu_y - \boldsymbol{\beta}_1'\boldsymbol{\mu}_x \text{ and } \boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}.$$

2.6.7.3 *Best Linear Prediction (BLP)*

Consider two random vectors, $\mathbf{x}\colon p \times 1$ and $\mathbf{y}\colon q \times 1$, with moments

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_y\\ \boldsymbol{\mu}_x\end{pmatrix} \text{ and } \mathrm{Var}\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} = \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx}\\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx}\end{pmatrix}.$$

Suppose that we wish to predict y after observing $\mathbf{x} = \ddot{\mathbf{x}}$. The joint density of $\mathbf{x}$ and $\mathbf{y}$ is not known, so the BP can not be used. Instead, we will find the BLP. The BLP minimizes

$$MSE(\hat{\mathbf{g}}) = \mathrm{E}\left[(\mathbf{y} - \hat{\mathbf{g}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \hat{\mathbf{g}})\right],$$

where $\boldsymbol{\Omega} > 0$ subject to

1. $\hat{\mathbf{g}}(\mathbf{x}) = \boldsymbol{\beta}_0 + \mathbf{B}_1'\mathbf{x}$, where $\boldsymbol{\beta}_0$ is a $q \times 1$ vector of constants and $\mathbf{B}_1$ is a $p \times q$ matrix of constants.
2. $\mathrm{E}(\hat{\mathbf{g}}) = \mathrm{E}(\mathbf{y})$.

Together, the two constraints imply that $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_y - \mathbf{B}_1'\boldsymbol{\mu}_x$. Thus, the BLP minimizes

$$MSE(\hat{\mathbf{g}}) = \mathrm{E}\left\{[(\mathbf{y} - \boldsymbol{\mu}_y) - \mathbf{B}_1'(\mathbf{x} - \boldsymbol{\mu}_x)]'\boldsymbol{\Omega}^{-1}[(\mathbf{y} - \boldsymbol{\mu}_y) - \mathbf{B}_1'(\mathbf{x} - \boldsymbol{\mu}_x)]\right\}.$$

Theorem 2.9 *The BLP is* $\hat{\mathbf{g}}(\mathbf{x}) = \boldsymbol{\beta}_0 + \mathbf{B}_1'\mathbf{x}$*, where* $\mathbf{B}_1 = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ *and* $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_y - \mathbf{B}_1'\boldsymbol{\mu}_x$*.*

□

Corollary. For $q = 1$, the BLP is $\hat{\mathbf{g}}(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}$, where $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ and $\beta_0 = \mu_y - \boldsymbol{\beta}_1'\boldsymbol{\mu}_x$.

Chapter 3

ESTIMATION OF B AND Σ FROM MVN

3.1 COMPLETE DATA

Consider the model $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, where $\text{vec}(\mathbf{U}) \sim \text{N}[\mathbf{0}, (\mathbf{\Sigma} \otimes \mathbf{\Omega})]$, $\mathbf{X}$ is an $n \times p$ known model matrix with rank $r \leq p$, $\mathbf{B}$ is an unknown $p \times d$ matrix of regression coefficients, $\mathbf{\Sigma} > 0$, and $\mathbf{\Omega}$ is a known $n \times n$ positive definite matrix.

3.1.1 Maximum Likelihood Estimator of B

Theorem 3.1 *A maximum likelihood estimator of* $\mathbf{B}$ *is any solution to the normal equations:*

$$\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{XB} = \mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y}.$$

One solution is

$$\widetilde{\mathbf{B}} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y},$$

where $(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-}$ *is any generalized inverse of* $\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}$*. If* $\mathbf{X}$ *has full column rank, then the estimator is unique and is given by* $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y}$*. Proof: in class.*

□

Corollary 1 If $\mathbf{\Omega} = \mathbf{I}_n$, then $\widetilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$.

Corollary 2: Partition $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix},$$

where $\mathbf{y}_i$ is $d \times 1$. If $\text{E}(\mathbf{y}_i) = \boldsymbol{\mu}$ for $i = 1, \ldots, n$ and $\mathbf{\Omega} = \mathbf{I}_n$, then the model simplifies to $\mathbf{y}_i \sim$ iid $\text{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ or $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, where $\mathbf{X} = \mathbf{1}_n$, and $\mathbf{B} = \boldsymbol{\mu}'$. The maximum likelihood estimator of $\mathbf{B}$ is $\widehat{\mathbf{B}} = n^{-1}\mathbf{1}_n'\mathbf{Y} = \overline{\mathbf{y}}'$. That is, $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{y}}$.

Corollary 3 If $\mathbf{C}'\,\text{vec}(\mathbf{B})$ is estimable, then the BLUE is $\mathbf{C}'\,\text{vec}(\widetilde{\mathbf{B}})$.

Corollary 4: If the linear function $\mathbf{C}_1'\mathbf{BC}_2$ is estimable, then

$$\text{Disp}(\mathbf{C}_1'\widetilde{\mathbf{B}}\mathbf{C}_2) = \mathbf{C}_2'\mathbf{\Sigma}\mathbf{C}_2 \otimes \mathbf{C}_1'(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-}\mathbf{C}_1.$$

It is readily shown that the linear function $\mathbf{C}_1'\mathbf{BC}_2$ is estimable iff $\mathbf{C}_1 \in \mathcal{R}(\mathbf{X}')$.

3.1.2 Maximum Likelihood Estimator of Σ

Theorem 3.2 *Let* $\mathbf{A}$: $q \times q$ *be a positive definite matrix and let* m *be a known scalar constant. Then*

$$\max_{\mathbf{\Phi}>0} \frac{\exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{\Phi}^{-1}\mathbf{A})\right\}}{|\mathbf{\Phi}|^{\frac{m}{2}}} = \frac{\exp\left\{-\frac{mq}{2}\right\} m^{\frac{mq}{2}}}{|\mathbf{A}|^{\frac{m}{2}}},$$

and the maximizer is

$$\widehat{\mathbf{\Phi}} = m^{-1}\mathbf{A}.$$

Proof: The Cholesky factorization of $\mathbf{\Phi}^{-1}$ *can be written as* $\mathbf{\Phi}^{-1} = \mathbf{\Gamma}\mathbf{\Gamma}'$. *Consider maximizing*

$$g(\mathbf{\Gamma}) = \frac{|\mathbf{A}|^{\frac{m}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{\Phi}^{-1}\mathbf{A})\right\}}{|\mathbf{\Phi}|^{\frac{m}{2}}} = |\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma}|^{\frac{m}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma})\right\}.$$

Note that

$$\ln\left[g(\mathbf{\Gamma})\right] = \frac{m}{2}\sum_{i=1}^{q}\ln(\lambda_i) - \frac{1}{2}\sum_{i=1}^{q}\lambda_i,$$

where λ_i *for* $i = 1, \ldots, q$ *are the eigenvalues of* $\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma}$. *Maximizing* $\ln\left[g(\mathbf{\Gamma})\right]$ *with respect to the* λ_i*s reveals that* $\mathbf{\Gamma}'\mathbf{A}\mathbf{\Gamma} = m\mathbf{I}$. *The result follows.*

□

Theorem 3.3 *Consider the model* $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}$ *where* $\operatorname{vec}(\mathbf{U}) \sim \mathrm{N}_{nd}[\mathbf{0}, (\mathbf{\Sigma} \otimes \mathbf{\Omega})]$. *Then, the MLE of* $\mathbf{\Sigma}$ *(for known* $\mathbf{\Omega}$*) is*

$$\widetilde{\mathbf{\Sigma}} = \frac{\mathbf{Y}'\mathbf{\Omega}^{-1}(\mathbf{I}_n - \mathbf{P})\mathbf{Y}}{n},$$

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{\Omega}^{-1}.$$

Note, if $\mathbf{\Omega} = \mathbf{I}$, *then* $\mathbf{P} = \operatorname{ppo}(\mathbf{X})$.

Proof: Write the likelihood function of $\mathbf{B}$ *and* $\mathbf{\Sigma}$ *given* $\mathbf{Y}$ *as*

$$L(\mathbf{B}, \mathbf{\Sigma}|\mathbf{Y}) = \frac{\exp\left\{-\frac{1}{2}\operatorname{tr}[\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{B})'\mathbf{\Omega}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{B})]\right\}}{|\mathbf{\Sigma}|^{\frac{n}{2}}|\mathbf{\Omega}|^{\frac{d}{2}}(2\pi)^{\frac{nd}{2}}}.$$

Maximize, first, with respect to $\mathbf{B}$. *Then apply the previous theorem.*

□

Corollary 1: $\mathrm{E}(\widetilde{\mathbf{\Sigma}}) = (n - r)\mathbf{\Sigma}/n$.

Corollary 2: $\mathrm{E}(\mathbf{S}) = \mathbf{\Sigma}$, where

$$\mathbf{S} = \frac{n\widetilde{\mathbf{\Sigma}}}{n - r} = \frac{\mathbf{Y}'\mathbf{\Omega}^{-1}(\mathbf{I}_n - \mathbf{P})\mathbf{Y}}{n - r}.$$

3.2 INCOMPLETE DATA: EM ALGORITHM

The EM algorithm is useful for computing MLEs when some data are missing.

3.2.1 References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society*, **B39**, 1–38.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley.

McLachlan, G. J., & Krishnan, T. (1997). *The EM Algorithm and Extensions*, New York: John Wiley.

Chapter 4

WISHART DISTRIBUTION

The Wishart distribution is a multivariate generalization of the gamma distribution.

Definition: Let $\mathbf{Y}$ be an $n \times d$ random matrix. Assume that $n \geq d$. Denote the i^{th} row of $\mathbf{Y}$ by $\mathbf{y}_i'$ and suppose that $\mathbf{y}_i$, for $i = 1, \ldots, n$, are independently distributed as $\mathbf{y}_i \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. That is,

$$\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}\left[\mathrm{vec}(\mathbf{M}), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)\right],$$

where $\mathbf{M} = \mathrm{E}(\mathbf{Y})$. Then, the $d \times d$ matrix $\mathbf{A} = \mathbf{Y}'\mathbf{Y}$ is said to have a d-dimensional Wishart distribution with n degrees of freedom, covariance matrix $\boldsymbol{\Sigma}$, and noncentrality matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}\mathbf{M}'\mathbf{M}$. The distribution of $\mathbf{A}$ is denoted by $\mathbf{A} \sim \mathrm{W}_d(n, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$. If $\boldsymbol{\Lambda} = \mathbf{0}$, then $\mathbf{A}$ is said to have a central Wishart distribution. A central Wishart distribution is denoted by $\mathrm{W}_d(n, \boldsymbol{\Sigma}, \mathbf{0})$ or, simply, $\mathrm{W}_d(n, \boldsymbol{\Sigma})$.

4.1 ANDERSON'S THEOREM

Anderson (1984, p. 245–249) gives a derivation of the central Wishart density. In the process, he proves a very useful result.

Theorem 4.1 *Let $\mathbf{Y}$ be an $n \times d$ matrix with density $f_Y(\mathbf{Y}'\mathbf{Y})$. That is, the density of $\mathbf{Y}$ depends on $\mathbf{Y}$ only through $\mathbf{A} = \mathbf{Y}'\mathbf{Y}$. Then, the density of $\mathbf{A}$ is*

$$f_A(\mathbf{A}) = \frac{|\mathbf{A}|^{\frac{n-d-1}{2}} f_Y(\mathbf{A}) \pi^{\frac{nd}{2}}}{\Gamma_d\left(\frac{n}{2}\right)},$$

where $\Gamma_d\left(\frac{n}{2}\right)$ is the multivariate gamma function:

$$\Gamma_d(t) = \pi^{\frac{d(d-1)}{4}} \prod_{i=1}^{d} \Gamma\left(t - \frac{i-1}{2}\right).$$

□

4.2 PROPERTIES OF THE WISHART DISTRIBUTION

Recall that if $\mathbf{y}_i \sim$ iid $\mathrm{N}_d(0, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n$, then

$$\mathbf{A} = \sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i' = \mathbf{Y}'\mathbf{Y} \sim \mathrm{W}_d(n, \boldsymbol{\Sigma}),$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}.$$

Theorem 4.2 *If* $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{\Sigma})$*, then the joint density of the distinct elements [that is the* $d(d+1)/2$ *elements in the upper or lower triangle] is*

$$f(\mathbf{A}) = \frac{|\mathbf{A}|^{\frac{n-d-1}{2}} \exp\{-\frac{1}{2}\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{A})\}}{2^{\frac{nd}{2}}|\mathbf{\Sigma}|^{\frac{n}{2}}\Gamma_d\left(\frac{n}{2}\right)},$$

for $\mathbf{A} > 0$ *and* $\mathbf{\Sigma} > 0$*.*

□

Theorem 4.3 *Suppose that* $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{\Sigma})$*. Let* $\mathbf{T}$ *be a symmetric matrix of constants. Then, the moment generating function of* $\mathbf{A}$ *is*

$$\mathrm{M}_{\mathbf{A}}(\mathbf{T}) = \mathrm{E}[\exp\{\operatorname{tr}(\mathbf{TA})\}] = |\mathbf{I}_d - 2\mathbf{T\Sigma}|^{-\frac{n}{2}}.$$

□

Theorem 4.4 *Suppose that* $\mathbf{Y}$ *is an* $n \times d$ *random matrix with distribution* $\operatorname{vec}(\mathbf{Y}) \sim \mathrm{N}_{nd}[\operatorname{vec}(\mathbf{M}), \mathbf{\Omega}]$ *where* $\mathbf{\Omega} > 0$ *and* $\mathrm{E}(\mathbf{Y}) = \mathbf{M}$*. Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix of constants. Then,* $\mathbf{Y}'\mathbf{AY} \sim \mathrm{W}_d(m, \mathbf{\Sigma}, \mathbf{\Lambda})$ *iff* $\mathbf{\Omega}(\mathbf{\Sigma}^{-1} \otimes \mathbf{A})$ *is idempotent. The parameters of the Wishart density are given by* $m = \operatorname{rank}(\mathbf{A})$*,*

$$\mathbf{\Sigma} = [\operatorname{tr}(\mathbf{A})]^{-1}\mathrm{T}_d[(\mathbf{I}_d \otimes \mathbf{A})\mathbf{\Omega}(\mathbf{I}_d \otimes \mathbf{A})],$$

and $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}\mathbf{M}'\mathbf{AM}$*.*

□

A proof of this remarkable result may be found in Appendix A of

Boik, R.J. (1988). The mixed model for multivariate repeated measures: validity conditions and an approximate test. *Psychometrika*, *53*, 469–486.

Corollary 1: Suppose that $\mathbf{Y}$ is an $n \times d$ random matrix with distribution $\operatorname{vec}(\mathbf{Y}) \sim \mathrm{N}_{nd}[\operatorname{vec}(\mathbf{M}), (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$, where $\mathbf{\Sigma} > 0$ and $\mathrm{E}(\mathbf{Y}) = \mathbf{M}$. Let $\mathbf{A}$ be an $n \times n$ symmetric matrix of constants. Then, $\mathbf{Y}'\mathbf{AY} \sim \mathrm{W}_d(m, \mathbf{\Sigma}, \mathbf{\Lambda})$ iff $\mathbf{A}$ is idempotent. The parameters of the Wishart density are given by $m = \operatorname{rank}(\mathbf{A})$ and $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}\mathbf{M}'\mathbf{AM}$.

Corollary 2: Consider the setup in Corollary 1 in which $\mathbf{M} = \mathbf{1}_n\boldsymbol{\mu}'$. That is, the rows of $\mathbf{Y}$ are iid $\mathrm{N}_d(\boldsymbol{\mu}, \mathbf{\Sigma})$. Define $\mathbf{H}$ by $\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$. Then, $\mathbf{Y}'\mathbf{HY} \sim \mathrm{W}_d(n-1, \mathbf{\Sigma}, \mathbf{0})$.

Corollary 3: Suppose $\mathbf{A}_i \sim$ ind $\mathrm{W}_d(n_i, \mathbf{\Sigma}, \mathbf{\Lambda}_i)$ for $i = 1, \ldots, k$. Then,

$$\sum_{i=1}^{k} \mathbf{A}_i \sim \mathrm{W}_d(n_\cdot, \mathbf{\Sigma}, \mathbf{\Lambda}_\cdot),$$

where $n_\cdot = \sum_{i=1}^k n_i$ and $\mathbf{\Lambda}_\cdot = \sum_{i=1}^k \mathbf{\Lambda}_i$. The proof of the Corollary consists of noting that $\mathbf{A}_i \sim \mathbf{Y}_i'\mathbf{Y}_i$ where $\operatorname{vec}(\mathbf{Y}_i) \sim \mathrm{N}_{n_i d}[\operatorname{vec}(\mathbf{M}_i), (\mathbf{\Sigma} \otimes \mathbf{I}_{n_i})]$ and $\mathbf{\Lambda}_i = \mathbf{\Sigma}^{-1}\mathbf{M}_i'\mathbf{M}_i$. Then, $\sum_{i=1}^k \mathbf{A}_i \sim \mathbf{Y}'\mathbf{Y}$ where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_k \end{pmatrix} \quad \text{and} \quad \operatorname{vec}(\mathbf{Y}) \sim \mathrm{N}[\operatorname{vec}(\mathbf{M}), (\mathbf{\Sigma} \otimes \mathbf{I}_{n_\cdot})].$$

Now use Corollary 1.

Theorem 4.5 *Suppose that* $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{I})$*. Write* $\mathbf{A}$ *in terms of its Cholesky decomposition:* $\mathbf{A} = \mathbf{R}'\mathbf{R}$ *where* $\mathbf{R}$ *is an upper triangular matrix having positive diagonal elements. The* $d(d+1)/2$ *elements in* $\mathbf{R}$ *are distributed, independently as*

$$r_{ij} \sim \begin{cases} \mathrm{N}(0,1), & \text{if } i < j; \\ +\sqrt{\chi^2(n-i+1)} & \text{if } i = j. \end{cases}$$

□

Corollary 1: Suppose that $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{I})$. Then,

$$|\mathbf{A}| = \prod_{i=1}^{d} r_{ii}^2 \sim \prod_{i=1}^{d} \chi^2(n-i+1),$$

where the χ^2's are mutually independent.

Theorem 4.6 *Suppose that* $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{\Sigma})$. *Let* $\mathbf{C}$ *be a* $d \times q$ *matrix of constants having rank-*q*, Then* $\mathbf{C}'\mathbf{A}\mathbf{C} \sim \mathrm{W}_q(n, \mathbf{C}'\mathbf{\Sigma}\mathbf{C})$.

□

Corollary 1: Suppose $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{\Sigma})$. Let $\mathbf{t}$ be any nonzero $d \times 1$ vector of constants. Then, $\mathbf{t}'\mathbf{A}\mathbf{t} \sim \sigma^2\chi^2(n)$ where $\sigma^2 = \mathbf{t}'\mathbf{\Sigma}\mathbf{t}$.

Corollary 2: Suppose that $\mathbf{A} \sim \mathrm{W}_d(n, \mathbf{\Sigma})$. Then,

$$|\mathbf{A}||\mathbf{\Sigma}|^{-1} \sim \prod_{i=1}^{d} \chi^2(n-i+1),$$

where the χ^2's are mutually independent.

Theorem 4.7 *Suppose that* $\mathbf{Y}$ *is an* $n \times d$ *random matrix with distribution* $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}_{nd}[\mathrm{vec}(\mathbf{M}), (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$, *where* $\mathbf{\Sigma} > 0$ *and* $\mathrm{E}(\mathbf{Y}) = \mathbf{M}$. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *symmetric matrices of constants. Then,* $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ *and* $\mathbf{Y}'\mathbf{B}\mathbf{Y}$ *are independent iff* $\mathbf{A}\mathbf{B} = \mathbf{0}$. *Proof: use univariate results.*

□

Theorem 4.8 *Suppose that* $\mathbf{Y}$ *is an* $n \times d$ *random matrix with distribution* $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}_{nd}[\mathrm{vec}(\mathbf{M}), (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$, *where* $\mathbf{\Sigma} > 0$ *and* $\mathrm{E}(\mathbf{Y}) = \mathbf{M}$. *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix of constants and let* $\mathbf{B}$ *be a* $p \times n$ *matrix of constants. Then,* $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ *and* $\mathbf{B}\mathbf{Y}$ *are independent iff* $\mathbf{B}\mathbf{A} = \mathbf{0}$. *Proof: use univariate results.*

□

Theorem 4.9 *Suppose that* $\mathbf{Y}$ *is an* $n \times d$ *random matrix with distribution* $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}_{nd}[\mathrm{vec}(\mathbf{M}), (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$, *where* $\mathbf{\Sigma} > 0$ *and* $\mathrm{E}(\mathbf{Y}) = \mathbf{M}$. *Write* d *as* $d = p + q$ *and partition* $\mathbf{Y}$ *as* $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 \end{pmatrix}$, *where* $\mathbf{Y}_1$ *is* $n \times p$ *and* $\mathbf{Y}_2$ *is* $n \times q$. *Partition* $\mathbf{M}$ *conformably and partition* $\mathbf{\Sigma}$ *as* $\mathbf{\Sigma} = \{\mathbf{\Sigma}_{ij}\}$ *for* $i, j = 1, 2$ *where* $\mathbf{\Sigma}_{11}$ *is* $p \times p$. *Then,*

$$\mathrm{vec}(\mathbf{Y}_1)|\mathbf{Y}_2 \sim \mathrm{N}_{np}[\mathrm{vec}(\mathbf{M}_{1\cdot 2}), (\mathbf{\Sigma}_{11\cdot 2} \otimes \mathbf{I}_n)],$$

where $\mathbf{M}_{1\cdot 2} = \mathbf{M}_1 + (\ddot{\mathbf{Y}}_2 - \mathbf{M}_2)\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ *and* $\mathbf{\Sigma}_{11\cdot 2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.

□

Corollary 1: Consider the setup in Theorem 4.9 in which $\mathbf{M} = \mathbf{1}_n\boldsymbol{\mu}'$. Partition $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

where $\boldsymbol{\mu}_1$ is $p \times 1$ and $\boldsymbol{\mu}_2$ is $q \times 1$. Then,

$$\mathrm{vec}(\mathbf{Y}_1)|\mathbf{Y}_2 \sim \mathrm{N}_{np}[\mathrm{vec}(\mathbf{M}_{1\cdot 2}), (\mathbf{\Sigma}_{11\cdot 2} \otimes \mathbf{I}_n)],$$

where $\mathbf{M}_{1\cdot 2} = \mathbf{1}_n\boldsymbol{\theta}' + \ddot{\mathbf{Y}}_2\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ and $\boldsymbol{\theta} = \boldsymbol{\mu}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$.

Corollary 2: Suppose that $\mathbf{A} \sim \mathbf{W}_{p+q}(n, \mathbf{\Sigma})$. Partition $\mathbf{A}$ and $\mathbf{\Sigma}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ and $\mathbf{\Sigma}_{11}$ are each $p \times p$. Then, the following hold.

1. $\mathbf{A}_{11\cdot 2} \sim \mathrm{W}_p(n-q, \mathbf{\Sigma}_{11\cdot 2})$ where $\mathbf{A}_{11\cdot 2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ and $\mathbf{\Sigma}_{11\cdot 2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.
2. $\mathbf{A}_{22\cdot 1} \sim \mathrm{W}_q(n-p, \mathbf{\Sigma}_{22\cdot 1})$ where $\mathbf{A}_{22\cdot 1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and $\mathbf{\Sigma}_{22\cdot 1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$.
3. $\mathbf{A}_{11\cdot 2}$ is independent of both $\mathbf{A}_{22}$ and $\mathbf{A}_{21}$.
4. $\mathbf{A}_{22\cdot 1}$ is independent of both $\mathbf{A}_{11}$ and $\mathbf{A}_{12}$.

Chapter 5

PRINCIPLES OF TEST CONSTRUCTION

5.1 LIKELIHOOD RATIO TESTS

Consider a random matrix $\mathbf{Y}$ having density $f(\mathbf{Y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters. If some parameters are in matrix form, then vec or vech them. The likelihood function is obtained by considering $f(\mathbf{Y}|\boldsymbol{\theta})$ as a function of the parameters given the data. The likelihood function for $\boldsymbol{\theta}$ given $\mathbf{Y}$ is written as $L(\boldsymbol{\theta}|\mathbf{Y})$.

Suppose that a test of $\mathrm{H}_0\colon \boldsymbol{\theta} \in \Omega_0$ against $\mathrm{H}_a\colon \boldsymbol{\theta} \in \Omega_a$ is desired. It is assumed that $\Omega_0 \cap \Omega_a = \emptyset$, otherwise the hypotheses may not be sensible. Define Ω by $\Omega = \Omega_0 \cup \Omega_a$. The likelihood ratio (LR) criterion is defined as

$$\Lambda(\mathbf{Y}) = \frac{L_0}{L_a},$$

where

$$L_0 = \sup_{\theta \in \Omega_0} L(\boldsymbol{\theta}|\mathbf{Y}) \quad \text{and} \quad L_a = \sup_{\theta \in \Omega} L(\boldsymbol{\theta}|\mathbf{Y}).$$

The null hypothesis is rejected for small values of the LR criterion. The above test also is the likelihood ratio test of $\mathrm{H}_0\colon \boldsymbol{\theta} \in \Omega_0$ against $\mathrm{H}_a\colon \boldsymbol{\theta} \in \Omega$.

Suppose that the dimension of the parameter space Ω_0 is r and that the dimension of the parameter space Ω is s. The dimension of a parameter space is equal to the number of functionally independent parameters which are free to vary. Then, under some fairly general regularity conditions and a true null hypothesis,

$$-2\,\ln(\Lambda) \xrightarrow{\text{dist}} \chi^2(s-r)$$

as $n \to \infty$. This is the kind of result which is proven in STAT 550.

5.2 UNION INTERSECTION TESTS

The union intersection (UI) procedure was developed by S.N. Roy in 1953. It provides an alternative way of constructing a multi-parameter test. Consider the same setup as for the LR test. If $s - r = 1$, then the LR and UI tests are identical. However, when $s - r \geq 2$, the LR and UI tests may differ. In general, for testing multi-parameter (composite) hypotheses, there is no uniformly most powerful test. Accordingly, it is sensible to consider the merits of a variety of tests.

To construct a UI test of $\mathrm{H}_0\colon \boldsymbol{\theta} \in \Omega_0$ against $\mathrm{H}_a\colon \boldsymbol{\theta} \in \Omega_a$, the null is rewritten as an intersection of one-parameter hypotheses and the alternative is rewritten as a union of one-parameter hypotheses. Let $g_i(\boldsymbol{\theta})$ for $i = 1, \ldots, K$ be a set of scalar valued functions of $\boldsymbol{\theta}$. Let $\Omega_{0,i}$ be the parameter space induced by the transformation from $\boldsymbol{\theta}$ to $g_i(\boldsymbol{\theta})$. That is,

$$\Omega_{0,i} = \{\tau | \tau = g_i(\boldsymbol{\theta}) \text{ for some } \boldsymbol{\theta} \in \Omega_0\}.$$

The parameter space $\Omega_{a,i}$ is defined similarly:

$$\Omega_{a,i} = \{\tau | \tau = g_i(\boldsymbol{\theta}) \text{ for some } \boldsymbol{\theta} \in \Omega_a\}.$$

The composite null and alternative hypotheses can be written as

$$\bigcap_{i=1}^{K} \mathrm{H}_{0,i} \colon g_i(\boldsymbol{\theta}) \in \Omega_{0,i} \quad \text{and} \quad \bigcup_{i=1}^{K} \mathrm{H}_{a,i} \colon g_i(\boldsymbol{\theta}) \in \Omega_{a,i}.$$

That is, for H_0 to be true, $g_i(\boldsymbol{\theta}) \in \Omega_{0,i}$ must be satisfied for <u>all</u> i and for H_a to be true, $g_i(\boldsymbol{\theta}) \in \Omega_{a,i}$ must be satisfied for <u>some</u> i. In practice, K may be infinite.

For example, suppose that we wish to test $\mathrm{H}_0 \colon \boldsymbol{\mu} = \mathbf{0}$ where $\boldsymbol{\mu}$ is a d-vector of population means. A finite union intersection set of hypotheses is obtained by defining g_i as $g_i(\boldsymbol{\mu}) = \mu_i$. In this case, the UI hypotheses are

$$\bigcap_{i=1}^{d} \mathrm{H}_{0,i} \colon \mu_i = 0 \quad \text{versus} \quad \bigcup_{i=1}^{d} \mathrm{H}_{a,i} \colon \mu_i \neq 0.$$

An infinite union intersection set of hypotheses is obtained by defining g_i as $g_i(\boldsymbol{\mu}) = \mathbf{t}_i' \boldsymbol{\mu}$, where $\mathbf{t}_i$ is a d-vector of coefficients. In this case, the UI hypotheses are

$$\mathbf{t}' \boldsymbol{\mu} = 0 \ \forall \ \mathbf{t} \quad \text{versus} \quad \mathbf{t}' \boldsymbol{\mu} \neq 0 \text{ for some } \mathbf{t}.$$

Let S_i be a test statistic for testing

$$\mathrm{H}_{0,i} \colon g_i(\boldsymbol{\theta}) \in \Omega_{0,i} \quad \text{versus} \quad \mathrm{H}_{a,i} \colon g_i(\boldsymbol{\theta}) \in \Omega_{a,i}.$$

In practice, S_i is typically the LR statistic for testing $\mathrm{H}_{0,i}$ against $\mathrm{H}_{a,i}$. Assume that the null is rejected if S_i is large. Then the UI test rejects the composite null if S is large, where

$$S = \sup_i S_i.$$

To perform a size α test, the null distribution of S is needed. In practice, it is usually easier to derive the UI test than it is to derive the null distribution of the UI test statistic. A more complete description of the UI principle can be found in Srivastava & Khatri *An Introduction to Multivariate Statistics*, 1979, p. 104–110.

Chapter 6

MULTIVARIATE TEST STATISTICS

In univariate linear models, the usual test statistics are functions of two independent sums of squares, SSE and SSH. Typically, SSE and SSH are independently distributed as

$$\frac{SSH}{\sigma^2} \sim \chi^2(\nu_1, \lambda) \quad \text{and} \quad \frac{SSE}{\sigma^2} \sim \chi^2(\nu_2, 0).$$

A size α test of $\mathrm{H}_0\colon \lambda = 0$ is given by the following: reject H_0 if $F \geq F^{1-\alpha}_{\nu_1,\nu_2}$ where

$$F = \left(\frac{SSH}{SSE}\right)\left(\frac{\nu_2}{\nu_1}\right),$$

and $F^{1-\alpha}_{\nu_1,\nu_2}$ is the $100(1-\alpha)$ percentile of the central F distribution with ν_1 and ν_2 degrees of freedom. An identical test is given by the following decision rule: reject H_0 if $B_1 \geq B(1-\alpha, \frac{\nu_1}{2}, \frac{\nu_2}{2})$ where

$$B_1 = \frac{SSH}{SSH + SSE},$$

and $B(1-\alpha, \frac{\nu_1}{2}, \frac{\nu_2}{2})$ is the $100(1-\alpha)$ percentile of the central Beta distribution with parameters $\frac{\nu_1}{2}$ and $\frac{\nu_2}{2}$. A third, identical, test is the following: reject H_0 if $B_2 \leq B(\alpha, \frac{\nu_2}{2}, \frac{\nu_1}{2})$ where

$$B_2 = \frac{SSE}{SSH + SSE},$$

and $B(\alpha, \frac{\nu_2}{2}, \frac{\nu_1}{2})$ is the 100α percentile of the central Beta distribution with parameters $\frac{\nu_2}{2}$ and $\frac{\nu_1}{2}$.

Many of the multivariate test statistics can be expressed as functions of two independent Wishart matrices, $\mathbf{E}$ and $\mathbf{H}$. Typically, $\mathbf{H}$ and $\mathbf{E}$ are independently distributed as

$$\mathbf{H} \sim \mathrm{W}_d(m_H, \mathbf{\Sigma}, \mathbf{\Lambda}) \quad \text{and} \quad \mathbf{E} \sim \mathrm{W}_d(m_E, \mathbf{\Sigma}, \mathbf{0}).$$

In some applications, m_H may be less than d, in which case, $\mathbf{H}$ will have a singular Wishart distribution. We assume that $m_E \geq d$ so that $\mathbf{E}$ has a nonsingular Wishart distribution.

We wish to test $\mathrm{H}_0\colon \mathbf{\Lambda} = \mathbf{0}$. Denote the ordered nonzero characteristic roots of $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$ by $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_s \geq 0$, where $s = \mathrm{rank}(\mathbf{H}) = \min(m_H, d)$. The usual multivariate test statistics are functions of the θ_i's. In the univariate case, $s = 1$, $B_1 = \theta_1$, $B_2 = 1 - \theta_1$, and

$$F = \frac{m_E\theta_1}{m_H(1-\theta_1)}.$$

6.1 WILKS'S LAMBDA

A size α test is given by the following: reject H_0 if $U < U(\alpha, d, m_H, m_E)$ where

$$U = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|},$$

and $U(\alpha, d, m_H, m_E)$ is the 100α null percentile of the U distribution. The U statistic is a multivariate generalization of the beta random variable, B_2. Note that

$$U = \prod_{i=1}^{s}(1-\theta_i),$$

where $s = \min(m_H, d)$.

Lemma Suppose that T and Z are scalar random variables having finite ranges. If $\mathrm{E}(T^i) = \mathrm{E}(Z^i)$ for $i = 0, 1, \ldots, \infty$, then T and Z have identical distributions.

Theorem 6.1 *The null distribution of U is*

$$U \sim \prod_{i=1}^{d} B_i, \quad \text{where} \quad B_i \overset{\text{ind}}{\sim} \mathrm{B}\left(\frac{m_E - i + 1}{2}, \frac{m_H}{2}\right).$$

Proof: Recall that if $T \sim \mathrm{B}(\alpha, \beta)$, then

$$\mathrm{E}(T^h) = \frac{\Gamma(\alpha+\beta)\,\Gamma(\alpha+h)}{\Gamma(\alpha)\,\Gamma(\alpha+\beta+h)}.$$

The proof consists of showing that

$$\mathrm{E}(U^h) = \prod_{i=1}^{d} \frac{\Gamma\left(\frac{m_E+2h+1-i}{2}\right)\Gamma\left(\frac{m_E+m_H+1-i}{2}\right)}{\Gamma\left(\frac{m_E+1-i}{2}\right)\Gamma\left(\frac{m_E+m_H+2h+1-i}{2}\right)},$$

and then using the lemma. Details are an exercise.

□

Theorem 6.2 *Suppose that $U_1 \sim U(d, m_H, m_E)$ and $U_2 \sim U(m_h, d, m_E + m_H - d)$. Then, $U_1 \sim U_2$. The proof consists of showing that $\mathrm{E}(U_1^h) = \mathrm{E}(U_2^h)$ for all $h \geq 0$.*

□

The most widely used approximation to the distribution of U is due to Rao (1951). Rao showed that, under H_0,

$$\frac{(ft-g)(1-U^{\frac{1}{t}})}{m_H\, d U^{\frac{1}{t}}} \dot{\sim} F_{d\,m_H,\, ft-g},$$

where

$$f = m_E - \frac{d - m_H + 1}{2}, \qquad g = \frac{d\,m_H - 2}{2},$$

and

$$t = \left(\frac{d^2 m_H^2 - 4}{d^2 + m_H^2 - 5}\right)^{\frac{1}{2}}.$$

If $d\,m_H = 2$, then t is set to 1. Rao's approximation is exact if $\min(d, m_H) \leq 2$.

An alternative expansion of the characteristic function yields a χ^2 approximation to the distribution of U. The χ^2 expansion is due to Box (1954). Using only the first term in the expansion:

$$f \ln(U) \dot{\sim} \chi^2(d\,m_H),$$

where

$$f = m_E - \frac{d - m_H + 1}{2}.$$

Table D13 in Seber (1984) gives percentiles of the U distribution in terms of the chi squared approximation. Table D13 gives correction factors, C_α, such that

$$\Pr\left[-f \ln(U) \geq C_\alpha \chi^2(1-\alpha, d\,m_H)\right] = \alpha.$$

To use D13, define M as $M = m_E - d + 1$ and define f as above. Note, from the table, that for $C_\alpha \to 1$ as $M \to \infty$. Table A.9 in Rencher (2002) gives lower percentiles of U.

6.2 PILLAI'S TRACE

Pillai's trace statistic is defined as

$$V^{(s)} = \operatorname{tr}\left[\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}\right],$$

which also can be written as

$$V^{(s)} = \sum_{i=1}^{s} \theta_i.$$

An accurate approximation to the null distribution of $V^{(s)}$ is

$$\left(\frac{2\nu_2+s+1}{2\nu_1+s+1}\right)\left(\frac{V^{(s)}}{s-V^{(s)}}\right) \;\dot\sim\; F\left[s(2\nu_1+s+1), s(2\nu_2+s+1)\right],$$

where

$$s = \min)d, m_H), \quad \nu_1 = \frac{|d-m_H|-1}{2} \quad \text{and} \quad \nu_2 = \frac{m_E-d-1}{2}.$$

Exact critical values are tabled in D16 in Seber (1984) and in Table A.11 in Rencher (2002).

6.3 LAWLEY-HOTELLING TRACE

The Lawley-Hotelling trace is defined as

$$T_g^2 = m_E \operatorname{tr}(\mathbf{E}^{-1}\mathbf{H}),$$

which can be written as

$$T_g^2 = m_E \sum_{i=1}^{s}\left(\frac{\theta_i}{1-\theta_i}\right).$$

The statistic also is called Hotelling's generalized T^2. An accurate approximation to the null distribution of T_g^2, due to McKeon (1974), is

$$\operatorname{tr}(\mathbf{E}^{-1}\mathbf{H})\left(\frac{b(m_E-d-1)}{d\,m_H(b-2)}\right) \;\dot\sim\; F_{d\,m_H,b},$$

where

$$b = 4+\frac{d\,m_H+2}{B-1} \quad \text{and} \quad B = \frac{(m_E+m_H-d-1)(m_E-1)}{(m_E-d-3)(m_E-d)}.$$

An alternative approximation (apparently less accurate but used by SAS) is

$$\operatorname{tr}(\mathbf{E}^{-1}\mathbf{H})\left(\frac{2(s\nu_2+1)}{s^2(2\nu_1+s+1)}\right) \;\dot\sim\; F[s(2\nu_1+s+1), 2(s\nu_2+1)],$$

where $s=\min(m_H,d)$, $\nu_1 = (|d-m_H|-1)/2$, and $\nu_2 = (m_E-d-1)/2$. Exact null percentage points for $(m_E/m_H)\operatorname{tr}(\mathbf{E}^{-1}\mathbf{H})$ are given in Table D15 of Seber (1984) and in Table A.12 in Rencher (2002).

Alternative approximations to the null and non-null distribution of T_g^2 can be found in

van der Merwe & Crowther (1984), "An approximation to the distribution of Hotelling's generalized T_0^2-statistic," *South African Statistical Journal*, *18*, 68–90.

6.4 ROY'S MAXIMUM ROOT

Roy's test statistic is θ_1, the maximum root of $(\mathbf{H}+\mathbf{E})^{-1}\mathbf{H}$. An equivalent test statistic is

$$\varphi_1 = \frac{\theta_1}{1-\theta_1}$$

which is the maximum root of $\mathbf{E}^{-1}\mathbf{H}$. Table D14 in Seber (1984) and Table A.10 in Rencher (2002) give exact percentiles for θ_1. In Table D14 of Seber (1984), the definitions $s=\min(d,m_H)$, $\nu_1=\frac{1}{2}(|d-m_H|-1)$, and $\nu_2=\frac{1}{2}(m_E-d-1)$ are used. In Table A.10 in Rencher (2002), the definitions $s=\min(d,m_H)$, $m=\frac{1}{2}(|d-m_H|-1)$, and $N=\frac{1}{2}(m_E-d-1)$ are used. Chart 9 and Tables 6–14 in Morrison (1990) give percentiles for θ_1. In Morrison, the definitions $s=\min(d,m_H)$, $m=\frac{1}{2}(|d-m_H|-1)$, and $n=\frac{1}{2}(m_E-d-1)$ are used.

SAS approximates the distribution of $\varphi_1(m_E-r-1)/r$ where $r=\max(m_H,d)$ by an F random variable. The p-value provided by SAS is a lower bound on the true p-value, so use caution!

Chapter 7

HOTELLING'S T^2

7.1 ONE SAMPLE SETTING

7.1.1 The Test Statistic and its Distribution

Consider the model

$$\mathbf{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \mathbf{U},$$

where $\mathbf{Y}$ is $n \times d$ and $\text{vec}(\mathbf{U}) \sim \text{N}[\mathbf{0}, (\boldsymbol{\Sigma} \otimes \mathbf{I})]$. A test of $\text{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is desired.

Theorem 7.1 *The LR test of* $\text{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ *versus* $\text{H}_a\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ *is to reject* H_0 *for large values of* T^2*, where*

$$T^2 = n(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu}_0);$$

$$\overline{\mathbf{y}} = n^{-1}\mathbf{Y}'\mathbf{1}_n; \quad \mathbf{S} = (n-1)^{-1}\mathbf{A}; \quad \mathbf{A} = \mathbf{Y}'[\mathbf{I}_n - \mathbf{X}_x]\mathbf{Y};$$

and $\mathbf{H}_x = \text{ppo}(\mathbf{1}_n)$.
Proof: in class

□

Note that $\mathbf{S}$ is an unbiased estimator of $\boldsymbol{\Sigma}$.

Theorem 7.2 *The test that rejects* $\text{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ *in favor of* $\text{H}_a\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ *for large values of* T^2 *also is a union intersection test.*
Proof: Note that H_0 *says that* $\mathbf{c}'\boldsymbol{\mu} = \mathbf{c}'\boldsymbol{\mu}_0$ *for all* $d \times 1$ *vectors* $\mathbf{c}$*. Also* H_a *says that* $\mathbf{c}'\boldsymbol{\mu} \neq \mathbf{c}'\boldsymbol{\mu}_0$ *for some* $d \times 1$ *vector* $\mathbf{c}$*. Note that for fixed* $\mathbf{c}$*, the distribution of* $\mathbf{c}'\overline{\mathbf{y}}$ *is* $\mathbf{c}'\overline{\mathbf{y}} \sim \text{N}(\mathbf{c}'\boldsymbol{\mu}, n^{-1}\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$*. Consider the usual* t *test for testing* $\text{H}_0\colon \mathbf{c}'\boldsymbol{\mu} = \mathbf{c}'\boldsymbol{\mu}_0$ *against* $\text{H}_a\colon \mathbf{c}'\boldsymbol{\mu} \neq \mathbf{c}'\boldsymbol{\mu}_0$*. This test rejects* $\text{H}_0\colon \mathbf{c}'\boldsymbol{\mu} = \mathbf{c}'\boldsymbol{\mu}_0$ *for large values of* $|t_{\mathbf{c}}|$*, where*

$$t_{\mathbf{c}} = \sqrt{n}\frac{\mathbf{c}'(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)}{\sqrt{\mathbf{c}'\mathbf{S}\mathbf{c}}}.$$

Now maximize $t_{\mathbf{c}}^2$ *over* $\mathbf{c}$*.*

□

To examine the large sample null distribution of T^2 under non-normality, a multivariate version of the central limit theorem is needed.

Theorem 7.3 (Multivariate CLT) *Let* $\mathbf{Y}\colon n \times d$ *be a random matrix having iid rows each with expectation* $\boldsymbol{\mu}$ *and variance* $\boldsymbol{\Sigma}$*. Thus,* $\text{E}(\mathbf{Y}) = \mathbf{1}_n\boldsymbol{\mu}'$ *and* $\text{disp}(\mathbf{Y}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$*. Let* $\overline{\mathbf{y}} = n^{-1}\mathbf{Y}'\mathbf{1}_n$*. Then, as* $n \to \infty$*,*

$$\sqrt{n}(\overline{\mathbf{y}} - \boldsymbol{\mu}) \xrightarrow{\text{dist}} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Proof: Extra Credit.

□

To obtain the limiting distribution of T^2, two additional results are useful. It can be shown that

$$\mathbf{S} = \boldsymbol{\Sigma} + O_p(n^{-\frac{1}{2}}) \text{ and } \mathbf{S}^{-1} = \boldsymbol{\Sigma}^{-1} + O_p(n^{-\frac{1}{2}}),$$

where O_p means "order in probability" (see Bishop, Fienberg and Holland, 1975), This means that $\sqrt{n}\|\mathbf{S} - \boldsymbol{\Sigma}\|$ and $\sqrt{n}\|\mathbf{S}^{-1} - \boldsymbol{\Sigma}^{-1}\|$ each are bounded in probability. It also implies that $\mathbf{S}$ converges in probability to $\boldsymbol{\Sigma}$ and that $\mathbf{S}^{-1}$ converges in probability to $\boldsymbol{\Sigma}^{-1}$. It follows that

$$T^2 = n(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu}_0) + O_p(n^{-\frac{1}{2}}).$$

The second term on the right-hand-side converges in probability to zero. Using the multivariate CLT, the first term converges in distribution to a χ^2 random variable having d degrees of freedom. Thus,

$$T^2 \xrightarrow{\text{dist}} \chi^2_d$$

even if multivariate normality is not satisfied.

To obtain the small sample distribution of T^2 under normality, we will use the following theorem.

Theorem 7.4 *Suppose that* $\mathbf{A}$ *and* $\mathbf{u}$ *are independently distributed as*

$$\mathbf{A} \sim \mathrm{W}_d(m, \boldsymbol{\Sigma}) \quad \text{and} \quad \mathbf{u} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then,

$$\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} \sim \frac{\chi^2_{d,\lambda}}{\chi^2_{m-d+1}},$$

where

$$\lambda = \frac{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2},$$

and the two chi squared random variables are independently distributed.

Proof: Let $\mathbf{z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{u}$ *and let* $\mathbf{V} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{A}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. *Then,*

$$\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} = \mathbf{z}'\mathbf{V}^{-1}\mathbf{z}.$$

Also $\mathbf{z}$ *and* $\mathbf{V}$ *are independently distributed as*

$$\mathbf{z} \sim \mathrm{N}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}, \mathbf{I}_d\right) \text{ and } \mathbf{V} \sim \mathrm{W}_d(m, \mathbf{I}_d).$$

Let $\mathbf{Q}$ *be a* $d \times d$ *orthogonal matrix whose first column is* $\mathbf{z}(\mathbf{z}'\mathbf{z})^{-\frac{1}{2}}$. *Conditional on* $\mathbf{z}$, *the matrix* $\mathbf{H} = \mathbf{Q}'\mathbf{V}\mathbf{Q} \sim \mathrm{W}_d(m, \mathbf{I}_d)$. *This distribution does not depend on* $\mathbf{z}$. *Accordingly, it can be concluded that* $\mathbf{H} \sim \mathrm{W}_d(m, \mathbf{I}_d)$, *unconditionally and that* $\mathbf{H} \perp\!\!\!\perp \mathbf{z}$. *Note that*

$$\mathbf{z}'\mathbf{V}^{-1}\mathbf{z} = \mathbf{z}'\mathbf{Q}\left(\mathbf{Q}'\mathbf{V}\mathbf{Q}\right)^{-1}\mathbf{Q}'\mathbf{z} = \sqrt{\mathbf{z}'\mathbf{z}}\,\mathbf{e}_1'\mathbf{H}^{-1}\mathbf{e}_1\sqrt{\mathbf{z}'\mathbf{z}} = \frac{\mathbf{z}'\mathbf{z}}{h_{11\cdot 2}},$$

where $\mathbf{e}_1$ *is the first column of* $\mathbf{I}_d$ *and* $h_{11\cdot 2} \sim \mathbf{W}_1(m - d + 1, 1)$. *That is* $h_{11\cdot 2} \sim \chi^2_{m-d+1}$. *Also,* $\mathbf{z}'\mathbf{z} \sim \chi^2_{d,\lambda}$, *where* $\lambda = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2$.

□

Theorem 7.5 *Under multivariate normality,*

$$\left(\frac{n-d}{d(n-1)}\right)T^2 \sim F_{d,n-d,\lambda},$$

where

$$\lambda = n\frac{(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)}{2}.$$

Proof: Let $m = n - 1$, $\mathbf{A} = (n-1)\mathbf{S}$, *and* $\mathbf{u} = \sqrt{n}(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)$. *Now use Theorem 7.4. Note that under* $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$, *the test statistic* $T^2 \times (n-d)/[d(n-1)]$ *has a central* F *distribution.*

□

7.1.2 Simultaneous Confidence Intervals

In this section, we will construct simultaneous confidence intervals for linear functions of $\boldsymbol{\mu}$. The pivotal quantity method will be used. A pivotal quantity is a function of the data and the unknown parameters. Most importantly, the distribution of the pivotal quantity does not depend on unknown parameters. If a pivotal quantity can be identified, then a confidence interval can sometimes be obtained directly from probability statements made about the pivotal quantity.

Consider the function

$$Q = n(\overline{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu}).$$

Technically, Q is not a statistic because it depends on the unknown parameter vector, $\boldsymbol{\mu}$. From prior work, it is known that

$$\left[\frac{n-d}{d(n-1)}\right] \times Q \sim F_{d,n-d,0}.$$

The distribution of Q does not depend on $\boldsymbol{\mu}$, so Q is a pivotal quantity.

Clearly, the following probability statement is true:

$$\Pr\left[Q \leq \frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}\right] = 1-\alpha.$$

Let $\mathbf{c}$ be a $d \times 1$ vector. Consider the following function of $\overline{\mathbf{y}}$, $\boldsymbol{\mu}$, and $\mathbf{c}$:

$$G(\mathbf{c}) = \frac{n[\mathbf{c}'(\overline{\mathbf{y}} - \boldsymbol{\mu})]^2}{\mathbf{c}'\mathbf{S}\mathbf{c}}.$$

It can be shown that

$$\max_{\mathbf{c}} G(\mathbf{c}) = Q.$$

Therefore,

$$\Pr\left[\max_{\mathbf{c}} G(\mathbf{c}) \leq \frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}\right] = 1-\alpha,$$

which implies that

$$\Pr\left[G(\mathbf{c}) \leq \frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d} \;\forall\; \mathbf{c}\right] = 1-\alpha$$

and that

$$\Pr\left[-\sqrt{\frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}} \leq \frac{\sqrt{n}\mathbf{c}'(\overline{\mathbf{y}} - \boldsymbol{\mu})}{\mathbf{c}'\mathbf{S}\mathbf{c}} \leq \sqrt{\frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}} \;\forall\; \mathbf{c}\right] = 1-\alpha.$$

The above probability statement also can be written as

$$\Pr\left[\mathbf{c}'\overline{\mathbf{y}} - \sqrt{\mathbf{c}'\mathbf{S}\mathbf{c} \times F^*} \leq \mathbf{c}'\boldsymbol{\mu} \leq \mathbf{c}'\overline{\mathbf{y}} + \sqrt{\mathbf{c}'\mathbf{S}\mathbf{c} \times F^*} \;\forall\; \mathbf{c}\right] = 1-\alpha,$$

where

$$F^* = \left[\frac{d(n-1)}{n(n-d)}\right] \times F^{1-\alpha}_{d,n-d}.$$

The above results are summarized in the following theorem.

Theorem 7.6 *Consider the model* $\mathbf{Y} \sim \mathrm{N}[\mathrm{vec}(\mathbf{1}_n\boldsymbol{\mu}'), \boldsymbol{\Sigma} \otimes \mathbf{I}_n]$. *The BLUE of* $\boldsymbol{\mu}$ *is* $\overline{\mathbf{y}}$ *which has distribution* $\overline{\mathbf{y}} \sim \mathrm{N}(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma})$. *Simultaneous confidence intervals for all linear functions,* $\mathbf{c}'\boldsymbol{\mu}$, *are given by*

$$\mathbf{c}'\overline{\mathbf{y}} \pm \sqrt{\mathbf{c}'\mathbf{S}\mathbf{c}\left[\frac{d(n-1)}{n(n-d)}\right]F^{1-\alpha}_{d,n-d}}.$$

With probability $1-\alpha$, *all of the above intervals capture the appropriate linear function of* $\boldsymbol{\mu}$.

□

7.2 VARIATIONS IN THE ONE SAMPLE SETTING

7.2.1 Testing that $\mathrm{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \boldsymbol{\theta}_0$

Let $\mathbf{M}'\boldsymbol{\mu} = \boldsymbol{\theta}$ where $\mathbf{M}$ is an a priori $d \times q$ matrix having rank-q. Suppose that a test of $\mathrm{H}_0\colon \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is desired. To construct a test, postmultiply the model by $\mathbf{M}$ to obtain

$$\mathbf{YM} = \mathbf{1}_n\boldsymbol{\mu}'\mathbf{M} + \mathbf{UM} = \mathbf{1}_n\boldsymbol{\theta}' + \mathbf{U}^*.$$

Note that $\mathrm{vec}(\mathbf{U}^*) \sim \mathrm{N}[\mathbf{0}, (\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} \otimes \mathbf{I}_n)]$. The test is constructed by substituting $\mathbf{M}'\overline{\mathbf{y}}$ for $\overline{\mathbf{y}}$, $\boldsymbol{\theta}_0$ for $\boldsymbol{\mu}_0$, and $\mathbf{M}'\mathbf{SM}$ for $\mathbf{S}$ in T^2. The resulting test statistic is

$$T^2 = n\,(\mathbf{M}'\overline{\mathbf{y}} - \boldsymbol{\theta}_0)'(\mathbf{M}'\mathbf{SM})^{-1}(\mathbf{M}'\overline{\mathbf{y}} - \boldsymbol{\theta}_0),$$

and is distributed as

$$\left(\frac{(n-q)}{q(n-1)}\right) T^2 \sim F_{q,n-q,\lambda},$$

where

$$\lambda = \frac{n\,(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{2}.$$

7.2.2 Testing $\mathrm{H}_0\colon \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}$.

Let $\mathbf{X}^*$ be an a priori $d \times k$ matrix with rank-k and let $\boldsymbol{\beta}$ be an unknown $k \times 1$ parameter vector. Define q by $q = \dim\,[\mathcal{N}(\mathbf{X}^{*\prime})]$ It follows that $q = d - k$. Let $\mathbf{M}$ be a $d \times q$ matrix whose columns form a basis set for $\mathcal{N}(\mathbf{X}^{*\prime})$. Thus, $\mathcal{R}(\mathbf{M}) = \mathcal{N}(\mathbf{X}^{*\prime})$ and, by the fundamental theorem of linear algebra, $\mathcal{N}(\mathbf{M}') = \mathcal{R}(\mathbf{X}^*)$. It follows that

$$\boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta} \Longleftrightarrow \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}.$$

Accordingly, the null $\mathrm{H}_0\colon \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}$ can be tested by testing $\mathrm{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$. Using Section 7.2.1, the test statistic is

$$T^2 = n\,\overline{\mathbf{y}}'\mathbf{M}(\mathbf{M}'\mathbf{SM})^{-1}\mathbf{M}'\overline{\mathbf{y}},$$

and

$$\left(\frac{n-q}{q(n-1)}\right) T^2 \sim F_{q,n-q,\lambda},$$

where

$$\lambda = \frac{n\,\boldsymbol{\mu}'\mathbf{M}(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}\mathbf{M}'\boldsymbol{\mu}}{2}.$$

7.2.3 Alternative Method for Deriving the Test of $\mathrm{H}_0\colon \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}$.

Theorem 7.7 *The statistic for testing* $\mathrm{H}_0\colon \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}$ *can be obtained by minimizing* $n\,(\overline{\mathbf{y}} - \mathbf{X}^*\boldsymbol{\beta})'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \mathbf{X}^*\boldsymbol{\beta})$ *with respect to* $\boldsymbol{\beta}$. *That is,*

$$\min_{\beta} n\,(\overline{\mathbf{y}} - \mathbf{X}^*\boldsymbol{\beta})'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \mathbf{X}^*\boldsymbol{\beta}) = n\,\overline{\mathbf{y}}'\mathbf{M}(\mathbf{M}'\mathbf{SM})^{-1}\mathbf{M}'\overline{\mathbf{y}}.$$

Proof: Using standard linear models results, it is readily shown that

$$\min_{\beta} n\,(\overline{\mathbf{y}} - \mathbf{X}^*\boldsymbol{\beta})'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \mathbf{X}^*\boldsymbol{\beta}) = n\,\overline{\mathbf{y}}'[\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{S}^{-1}\mathbf{X}^*)^{-}\mathbf{X}^{*\prime}\mathbf{S}^{-1}]\overline{\mathbf{y}}.$$

To complete the proof, it must be shown that

$$\mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{S}^{-1}\mathbf{X}^*)^{-}\mathbf{X}^{*\prime}\mathbf{S}^{-1} = \mathbf{M}(\mathbf{M}'\mathbf{SM})^{-1}\mathbf{M}',$$

where the columns of $\mathbf{M}\colon d \times q$ *form a basis set for* $\mathcal{N}(\mathbf{X}^{*\prime})$. *To verify the above equality, let* $\mathbf{S}^{-\frac{1}{2}}$ *be the symmetric square root of* $\mathbf{S}$. *Show that* $\mathcal{R}(\mathbf{S}^{\frac{1}{2}}\mathbf{M}) = \mathcal{N}(\mathbf{X}^{*\prime}\mathbf{S}^{-\frac{1}{2}})$. *It then follows that* $\mathrm{ppo}(\mathbf{S}^{\frac{1}{2}}\mathbf{M}) = \mathbf{I} - \mathrm{ppo}(\mathbf{S}^{-\frac{1}{2}}\mathbf{X}^*)$. *Can you fill in the details concerning this line of reasoning? For help, see Khatri's Lemma in the 505 notes.*

□

7.2.4 Roy's Step Down Tests

This section is optional. It explains the theory underlying the tests for additional information described in Rencher (2002), Section 5.8.

Suppose, as above, we wish to test $\mathrm{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\delta}$. The change in notation from $\boldsymbol{\mu}_0$ to $\boldsymbol{\delta}$ is for convenience; i.e., it is easier to write $\boldsymbol{\delta}_i$ than to write $\boldsymbol{\mu}_{0,i}$. Suppose that we wish to place more emphasis on the first p elements of $\boldsymbol{\mu}$. We may, for example, believe that if $\boldsymbol{\mu}$ differs from $\boldsymbol{\delta}$, it will be the first p components which contribute most to the difference. Partition $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

where $\boldsymbol{\mu}_1$ is $p \times 1$, $\boldsymbol{\mu}_2$ is $q \times 1$, and $p + q = d$. Partition $\mathbf{Y}$ and $\boldsymbol{\delta}$ conformably:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix},$$

and

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix}.$$

As before, assume that $\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}[\mathrm{vec}(\mathbf{1}_n\boldsymbol{\mu}'), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)]$. Note that

$$\boldsymbol{\mu} = \boldsymbol{\delta} \Leftrightarrow \begin{cases} \mathrm{H}_1\colon \boldsymbol{\mu}_1 = \boldsymbol{\delta}_1 \\ \mathrm{H}_2\colon \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1 = \boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1 \end{cases} \quad \text{and}$$

are <u>both</u> true. Therefore, a test of H_0 can be constructed by testing H_1 and H_2. It is sometimes more convenient to write H_2 as

$$\mathrm{H}_2\colon \boldsymbol{\mu}_{2\cdot1} = \boldsymbol{\delta}_{2\cdot1},$$

where $\boldsymbol{\delta}_{2\cdot1} = \boldsymbol{\delta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\delta}_1$ and $\boldsymbol{\mu}_{2\cdot1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$.

Hotelling's T^2 statistic for testing H_1 is

$$T_1^2 = n\,(\overline{\mathbf{y}}_1 - \boldsymbol{\delta}_1)'\mathbf{S}_{11}^{-1}(\overline{\mathbf{y}}_1 - \boldsymbol{\delta}_1),$$

and is distributed as

$$\left(\frac{n-p}{p(n-1)}\right) T_1^2 \sim F_{p,n-p,\lambda},$$

where

$$\lambda = \frac{n\,(\boldsymbol{\mu}_1 - \boldsymbol{\delta}_1)'\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\delta}_1)}{2}.$$

By analogy, we might expect to test H_2 by using the statistic

$$T_2^2 = n\,(\overline{\mathbf{y}}_{2\cdot1} - \widehat{\boldsymbol{\delta}}_{2\cdot1})'\mathbf{S}_{22\cdot1}^{-1}(\overline{\mathbf{y}}_{2\cdot1} - \widehat{\boldsymbol{\delta}}_{2\cdot1}),$$

where

$$\overline{\mathbf{y}}_{2\cdot1} = \overline{\mathbf{y}}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\overline{\mathbf{y}}_1,$$
$$\widehat{\boldsymbol{\delta}}_{2\cdot1} = \boldsymbol{\delta}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\boldsymbol{\delta}_1,$$

and

$$\mathbf{S}_{22\cdot1} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}.$$

Note, $\boldsymbol{\delta}_{2\cdot1}$ depends on $\boldsymbol{\Sigma}$ so it must be estimated.

Theorem 7.8 *Conditional on $\mathbf{Y}_1$, the distribution of T_2^2 is proportional to an F. In particular,*

$$\left(\frac{n-d}{(n-1)q}\right) \frac{T_2^2}{1 + (n-1)^{-1}T_1^2} \,\Big|\, \mathbf{Y}_1 \sim F_{q,n-d,\lambda},$$

where

$$\lambda = \frac{(\boldsymbol{\mu}_{2\cdot1} - \boldsymbol{\delta}_{2\cdot1})'\boldsymbol{\Sigma}_{22\cdot1}^{-1}(\boldsymbol{\mu}_{2\cdot1} - \boldsymbol{\delta}_{2\cdot1})}{2[1 + (n-1)^{-1}T_1^2]}.$$

Outline of proof: First establish that

$$\text{vec}(\mathbf{Y}_2)|\mathbf{Y}_1 \sim \text{N}\left\{\text{vec}\left[\mathbf{1}_n\boldsymbol{\mu}_2' + (\mathbf{Y}_1 - \mathbf{1}_n\boldsymbol{\mu}_1')\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right], \boldsymbol{\Sigma}_{22\cdot1} \otimes \mathbf{I}_n\right\},$$

and

$$\overline{\mathbf{y}}_{2\cdot1} - \widehat{\boldsymbol{\delta}}_{2\cdot1} = \mathbf{Y}_2'\left[\mathbf{1}_n n^{-1} - (\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}_1\mathbf{A}_{11}^{-1}(\overline{\mathbf{y}}_1 - \boldsymbol{\delta}_1)\right] - \boldsymbol{\delta}_2,$$

where $\mathbf{H}_x = \text{ppo}(\mathbf{1}_n)$*. Use the above two results to establish*

$$\frac{\sqrt{n}(\overline{\mathbf{y}}_{2\cdot1} - \widehat{\boldsymbol{\delta}}_{2\cdot1})}{\sqrt{1 + (n-1)^{-1}T_1^2}}\Bigg|\mathbf{Y}_1 \sim \text{N}\left[\frac{\sqrt{n}(\boldsymbol{\mu}_{2\cdot1} - \boldsymbol{\delta}_{2\cdot1})}{\sqrt{1 + (n-1)^{-1}T_1^2}}, \boldsymbol{\Sigma}_{22\cdot1}\right].$$

Use the above, along with the result $\mathbf{A}_{22\cdot1} \sim \text{W}_q(n-1-p, \boldsymbol{\Sigma}_{22\cdot1})$ *to finish the proof.*

□

Corollary 1: If H_2 is true then

$$\left(\frac{n-d}{(n-1)q}\right)\left(\frac{T_2^2}{1 + (n-1)^{-1}T_1^2}\right) \sim F_{q,n-d},$$

unconditionally and

$$\left(\frac{n-d}{(n-1)q}\right)\left(\frac{T_2^2}{1 + (n-1)^{-1}T_1^2}\right) \perp\!\!\!\perp T_1^2.$$

Corollary 2: Suppose that a size α_1 test of H_1 is conducted and that a size α_2 test of H_2 is conducted. Then the test size for the two tests simultaneously (i.e., the test of H_0) is $1 - (1-\alpha_1)(1-\alpha_2)$.

Note 1: $T_2^2 = T^2 - T_1^2$ where $T^2 = n\,(\overline{\mathbf{y}} - \boldsymbol{\delta})'\mathbf{S}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\delta})$. This result can be established by expressing $\mathbf{S}$ as a partitioned matrix and using the expression for the inverse of a partitioned matrix.

Note 2: Roy's step-down tests can be generalized from 2 steps to d steps.

7.2.5 One Sample Profile Analysis

Consider the usual one sample multivariate model:

$$\text{vec}(\mathbf{Y}) \sim \text{N}[\text{vec}(\mathbf{1}_n\boldsymbol{\mu}'), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)].$$

Denote the i^{th} row of $\mathbf{Y}$ by $\mathbf{y}_i'$. In this section, the d responses in $\mathbf{y}_i$ are assumed to represent repeated measures on the same dependent variable. The line graph of the mean response as a function of time (measurement period) is called a profile. The analysis of a single profile usually focuses on answering two questions:

1. Location: What is the overall level of the profile? The profile location is given by $\mu_\cdot = d^{-1}\mathbf{1}_d'\boldsymbol{\mu}$. A confidence interval for $\mu_\cdot$ and/or a hypothesis test of $\text{H}_0\colon \mu_\cdot = \mu_{\cdot 0}$ may be desired.

2. Shape: How does the profile vary as a function of time? A confidence interval for a contrast among the elements of $\boldsymbol{\mu}$ (i.e., $\boldsymbol{\ell}'\boldsymbol{\mu}$, where $\boldsymbol{\ell}'\mathbf{1}_d = 0$) or a test of $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$ may be desired, where $\mathbf{M}$ is $d \times (d-1)$, satisfies $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$ and has rank $d-1$.

Each of these questions can be answered using one sample techniques that are related to Hotelling's T^2. The vector of profile means is $\boldsymbol{\mu}$ and is estimated by $\overline{\mathbf{y}}$:

$$\overline{\mathbf{y}} = \mathbf{Y}'\mathbf{1}_n n^{-1},$$

which has distribution

$$\overline{\mathbf{y}} \sim \text{N}(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}).$$

The profile location (average level) is $\mu_\cdot = \boldsymbol{\mu}'\mathbf{1}_d d^{-1}$. The corresponding MLE is $\overline{y}_\cdot = \overline{\mathbf{y}}'\mathbf{1}_d d^{-1}$, which is distributed as $\overline{y}_\cdot \sim \text{N}[\mu_\cdot, (nd^2)^{-1}\mathbf{1}_d'\boldsymbol{\Sigma}\mathbf{1}_d]$.

Theorem 7.9 (Location) *Confidence intervals and hypothesis tests concerning the profile location can be based on the following pivotal quantity:*

$$\frac{d\sqrt{n}(\overline{y}_{\cdot} - \mu_{\cdot})}{\sqrt{\mathbf{1}_d'\mathbf{S}\mathbf{1}_d}} \sim t_{n-1,0},$$

where $\mathbf{S} = (n-1)^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}$*, and* $\mathbf{H}_x = \text{ppo}(\mathbf{1}_n)$*. In particular,* $\text{H}_0\colon \mu_{\cdot} = \mu_{\cdot 0}$ *is rejected in favor of* $\text{H}_a\colon \mu_{\cdot} \neq \mu_{\cdot 0}$ *if*

$$\left|\frac{d\sqrt{n}(\overline{y}_{\cdot} - \mu_{\cdot 0})}{\sqrt{\mathbf{1}_d'\mathbf{S}\mathbf{1}_d}}\right| \geq t_{n-1}^{1-\alpha/2}.$$

Also, a $100(1-\alpha)\%$ *confidence interval for* $\mu_{\cdot}$ *can be constructed as*

$$\overline{y}_{\cdot} \pm t_{n-1}^{1-\alpha/2}\, d^{-1}\sqrt{n^{-1}\mathbf{1}_d'\mathbf{S}\mathbf{1}_d}.$$

Proof: Let $\mathbf{A} = \mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}$*; let* $SSE = \mathbf{1}_d'\mathbf{A}\mathbf{1}_d/d^2$ *and let* $\sigma^2 = \mathbf{1}_d'\mathbf{\Sigma}\mathbf{1}_d/d^2$*. Then, from Corollary 1 of Theorem 4.6,* $SSE\,/\sigma^2 \sim \chi^2_{n-1}$*. Also,* $\sqrt{n}(\overline{y}_{\cdot} - \mu_{\cdot}) \sim \text{N}(0, \sigma^2)$ *and* $\overline{y}_{\cdot}$ *is independent of SSE. Accordingly*

$$\frac{\sqrt{n}\dfrac{\overline{y}_{\cdot} - \mu_{\cdot}}{\sigma}}{\sqrt{\dfrac{SSE}{(n-1)\sigma^2}}} \sim t_{n-1,0}.$$

□

Questions concerning changes in the expected response over time (profile shape) can be answered by examining contrasts among the d time periods. Let $\mathbf{M}$ be a $d \times (d-1)$ matrix of contrast coefficients with rank $d-1$. For $\mathbf{M}$ to consist of contrast coefficients, $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$ must be satisfied. The vector of profile contrasts is $\mathbf{M}'\boldsymbol{\mu}$ and is estimated by $\mathbf{M}'\overline{\mathbf{y}}$ which has distribution

$$\mathbf{M}'\overline{\mathbf{y}} \sim \text{N}(\mathbf{M}'\boldsymbol{\mu}, n^{-1}\mathbf{M}'\mathbf{\Sigma}\mathbf{M}).$$

Theorem 7.10 (Shape) *Confidence intervals and hypothesis tests concerning profile shape can be based on the following pivotal quantity:*

$$\left(\frac{n-d+1}{(n-1)(d-1)}\right)(\overline{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}'(\overline{\mathbf{y}} - \boldsymbol{\mu}) \sim F_{d-1,n-d+1,0}.$$

In particular, to test $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \boldsymbol{\delta}$*, use*

$$\left(\frac{n-d+1}{(n-1)(d-1)}\right)T^2 \sim F_{d-1,n-d+1,\lambda},$$

where

$$T^2 = n(\mathbf{M}'\overline{\mathbf{y}} - \boldsymbol{\delta})'(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}(\mathbf{M}'\overline{\mathbf{y}} - \boldsymbol{\delta}) \text{ and } \lambda = \frac{n\,(\mathbf{M}'\boldsymbol{\mu} - \boldsymbol{\delta})'(\mathbf{M}'\mathbf{\Sigma}\mathbf{M})^{-1}(\mathbf{M}'\boldsymbol{\mu} - \boldsymbol{\delta})}{2}.$$

To obtain simultaneous confidence intervals on linear functions $\boldsymbol{\ell}'\boldsymbol{\mu}$*, where* $\boldsymbol{\ell}'\mathbf{1}_d = \mathbf{0}$*, use*

$$\boldsymbol{\ell}'\overline{\mathbf{y}} \pm \sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}\left[\frac{(d-1)(n-1)}{n(n-d+1)}\right]F_{d-1,n-d+1}^{1-\alpha}}.$$

With probability $1-\alpha$*, all of the above intervals capture the appropriate linear function of* $\boldsymbol{\mu}$*. Proof: Use Theorems 7.4 and 7.6.*

□

7.3 TWO SAMPLE SETTING

7.3.1 The Linear Model

The cell means model for the multivariate one-way classification with two groups is

$$\text{vec}(\mathbf{Y}) \sim \text{N}[\text{vec}(\mathbf{XB}), (\mathbf{\Sigma} \otimes \mathbf{I}_N)],$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix},$$

$\mathbf{Y}_1$ is $n_1 \times d$, $\mathbf{Y}_2$ is $n_2 \times d$, $n_1 + n_2 = N$,

$$\mathbf{X} = \bigoplus_{i=1}^{2} \mathbf{1}_{n_i} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{pmatrix}, \text{ and } \mathbf{B} = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \end{pmatrix}.$$

The model also can be written as

$$\text{vec}(\mathbf{Y}_i) \sim \text{ ind N}[\text{vec}(\mathbf{1}_{n_i}\boldsymbol{\mu}_i'), \mathbf{\Sigma} \otimes \mathbf{I}_{n_i}],$$

for $i = 1, 2$.

7.3.2 Two Sample Hotelling's T^2

Suppose that a test of $\text{H}_0\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against $\text{H}_a\colon \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ is desired. The hypotheses also can be written as $\text{H}_0\colon \text{E}(\mathbf{Y}) = \mathbf{X}_0\mathbf{B}_0$ versus $\text{H}_a\colon \text{E}(\mathbf{Y}) = \mathbf{XB}$, where $\mathbf{X}_0 = \mathbf{1}_N$, $\mathbf{B}_0 = \boldsymbol{\mu}'$, and the remaining terms are defined above.

Theorem 7.11 (Two-Sample T^2) *: The LR and UI tests of* $\text{H}_0\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ *versus* $\text{H}_a\colon \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ *are identical. The test is to reject* H_0 *for large values of*

$$T^2 = \left(\frac{n_1 n_2}{N}\right)(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{S}^{-1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2),$$

where

$$\mathbf{S} = \frac{\mathbf{Y}'(\mathbf{I}_N - \mathbf{H}_x)\mathbf{Y}}{N - r} = \frac{\mathbf{Y}'(\mathbf{I}_N - \mathbf{H}_x)\mathbf{Y}}{N - 2},$$

where $\mathbf{H}_x = \text{ppo}(\mathbf{X})$*, and* $r = \text{rank}(\mathbf{X}) = 2$.
Sketch of LR proof: Let $\mathbf{H}_{0x} = \text{ppo}(\mathbf{X}_0) = \text{ppo}(\mathbf{1}_N)$*. Then, under* H_0*, the MLE of* $\mathbf{\Sigma}$ *is*

$$\widehat{\mathbf{\Sigma}}_0 = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{x0})\mathbf{Y}}{N} = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y} + \mathbf{Y}'(\mathbf{H}_x - \mathbf{H}_{x0})\mathbf{Y}}{N}.$$

Note

$$\mathbf{H}_x - \mathbf{H}_{x0} = \text{ppo}\left[(\mathbf{I} - \mathbf{X}_{x0})\mathbf{X}\right] = \text{ppo}\begin{pmatrix} \frac{n_2}{N}\mathbf{1}_{n_1} & -\frac{n_2}{N}\mathbf{1}_{n_1} \\ -\frac{n_1}{N}\mathbf{1}_{n_2} & \frac{n_1}{N}\mathbf{1}_{n_2} \end{pmatrix}$$

$$= \text{ppo}\begin{pmatrix} n_2\mathbf{1}_{n_1} \\ -n_1\mathbf{1}_{n_2} \end{pmatrix} = \frac{1}{N}\begin{pmatrix} \frac{n_2}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}' & -\mathbf{1}_{n_1}\mathbf{1}_{n_2}' \\ -\mathbf{1}_{n_2}\mathbf{1}_{n_1}' & \frac{n_1}{n_2}\mathbf{1}_{n_2}\mathbf{1}_{n_2}' \end{pmatrix}.$$

Using the above expression, it is readily shown that

$$\mathbf{Y}'(\mathbf{H}_x - \mathbf{H}_{x0})\mathbf{Y} = \left(\frac{n_1 n_2}{N}\right)(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'.$$

The remainder of the proof follows the proof for Hotelling's one sample test.

□

Theorem 7.12 (Distribution of Two Sample T^2) *: The distribution of* T^2 *in Theorem 7.11 is the following:*

$$\left(\frac{N - d - 1}{d(N - 2)}\right)T^2 \sim F_{d,N-d-1,\lambda},$$

where

$$\lambda = \left(\frac{n_1 n_2}{N}\right)\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{2}.$$

Proof: Use Theorem 7.4.

□

7.3.3 Two Sample Profile Analysis

Consider the same setup as the one sample profile analysis problem except that data from two independent samples have been obtained. Each of the corresponding two populations can be characterized by a profile of means: $\boldsymbol{\mu}_i$ for $i = 1, 2$. A typical two sample profile analysis consists of answering three questions:

1. Location: Do the two profiles have the same average level? The null hypothesis is $\mathrm{H}_0\colon \mu_{1\cdot} = \mu_{2\cdot}$, where $\mu_{i\cdot} = \boldsymbol{\mu}_i'\mathbf{1}_d d^{-1}$. The null hypothesis states that the two profiles have the same location (average level).

2. Parallel: Are the two profiles parallel; i.e., do the profiles have the same shape? This question also can be asked as — is there an interaction between groups and time. The null hypothesis is $\mathrm{H}_0\colon \mathbf{M}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ where $\mathbf{M}$ is $d \times (d-1)$, has rank $d-1$, and satisfies $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$. The null hypothesis states that the two profiles consist of parallel line segments.

3. Shape of Average Profile: Does the average profile (averaged over the two groups) vary as a function of time? This question is concerned with the shape of the average profile. The null hypothesis is $\mathrm{H}_0\colon \mathbf{M}'\boldsymbol{\mu}_{\cdot} = \mathbf{0}$ where $\mathbf{M}$ is described in (b) and $\boldsymbol{\mu}_{\cdot}$ is a weighted average of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$). The null hypothesis states that the average profile is a line with slope equal to zero.

Theorem 7.13 *[Difference in Location] Confidence intervals and hypothesis tests concerning differences in the location of the two profiles can be based on the following pivotal quantity:*

$$d\sqrt{\frac{n_1 n_2}{N}}\ \frac{(\overline{y}_{1\cdot} - \overline{y}_{2\cdot}) - (\mu_{1\cdot} - \mu_{2\cdot})}{\sqrt{\mathbf{1}_d'\mathbf{S}\mathbf{1}_d}} \sim t_{N-2,0},$$

where $\mathbf{S} = (N-2)^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}$, $\mathbf{H}_x = \mathrm{ppo}(\mathbf{X})$, *and* $\mathbf{X} = \mathbf{1}_{n_1} \oplus \mathbf{1}_{n_2}$. *In particular,* $\mathrm{H}_0\colon \mu_{1\cdot} = \mu_{2\cdot}$ *is rejected for large* $|t|$, *where*

$$t = d\sqrt{\frac{n_1 n_2}{N}}\ \frac{\overline{y}_{1\cdot} - \overline{y}_{2\cdot}}{\sqrt{\mathbf{1}_d'\mathbf{S}\mathbf{1}_d}}.$$

The distribution of t *is*

$$t \sim t_{N-2,\lambda},\ \text{where } \lambda = d\sqrt{\frac{n_1 n_2}{N}}\ \frac{\mu_{1\cdot} - \mu_{2\cdot}}{\sqrt{2}\sqrt{\mathbf{1}_d'\boldsymbol{\Sigma}\mathbf{1}_d}}.$$

A $100(1-\alpha)\%$ *Also, a confidence interval for* $\mu_{1\cdot} - \mu_{2\cdot}$ *can be constructed as*

$$\overline{y}_{1\cdot} - \overline{y}_{2\cdot} \pm t_{N-2}^{1-\alpha/2}\sqrt{\frac{N\mathbf{1}_d'\mathbf{S}\mathbf{1}_d}{d^2 n_1 n_2}}.$$

□

Theorem 7.14 (Interaction — Shape Differences) *Confidence intervals and hypothesis tests concerning interaction between groups and time can be based on the following pivotal quantity:*

$$\left(\frac{(N-d)n_1 n_2}{N(N-2)(d-1)}\right)[(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]'\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}'[(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$

$$\sim F_{d-1,N-d,0},$$

where $\mathbf{S}$ *is given in Theorem 7.13. In particular,* $\mathrm{H}_0\colon \mathbf{M}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ *is rejected for large* T^2, *where*

$$T^2 = \left(\frac{n_1 n_2}{N}\right)(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)'\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}'(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2).$$

The distribution of T^2 *is*

$$\left(\frac{N-d}{(N-2)(d-1)}\right)T^2 \sim F_{d-1,N-d,\lambda},$$

where

$$\lambda = \left(\frac{n_1 n_2}{N}\right)\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{M}(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}\mathbf{M}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{2}.$$

Simultaneous confidence intervals can be obtained from

$$\Pr[\boldsymbol{\ell}'(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) - k \leq \boldsymbol{\ell}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leq \boldsymbol{\ell}'(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) + k \;\forall\; \boldsymbol{\ell} \in \mathcal{R}(\mathbf{M})] = 1 - \alpha,$$

where $k = \sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}\, F^*}$, *and*

$$F^* = \left(\frac{N(d-1)(N-2)}{n_1 n_2 (N-d)}\right) F^{1-\alpha}_{d-1,N-d}.$$

Proof: Use Theorem 7.4. □

There are two approaches to making inferences about the average profile. In approach 1, it is assumed that the profiles are parallel. That is $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + \mathbf{1}_d k$, where k is a scalar constant. An average profile can be obtained as a weighted average of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Let w be a number in $[0, 1]$. Then

$$\boldsymbol{\mu}_\cdot = w\boldsymbol{\mu}_1 + (1-w)\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + (1-w)\mathbf{1}_d k.$$

The choice of w influences the overall level, but not the shape of the average profile. The sample estimator is

$$\overline{\mathbf{y}}_\cdot = w\overline{\mathbf{y}}_1 + (1-w)\overline{\mathbf{y}}_2.$$

To choose a value for w, it is sensible to use the value that minimizes $\mathrm{Var}(\overline{\mathbf{y}}_\cdot)$. It is readily shown that the minimizer of $\mathrm{Var}(\overline{\mathbf{y}}_\cdot)$ with respect to w is $w = n_1/N$, where $N = n_1 + n_2$. Accordingly, $\boldsymbol{\mu}_\cdot$ is defined as

$$\boldsymbol{\mu}_\cdot = \frac{1}{N}(n_1\boldsymbol{\mu}_1 + n_2\boldsymbol{\mu}_2).$$

The corresponding estimator and its variance are

$$\overline{\mathbf{y}}_\cdot = \frac{n_1\overline{\mathbf{y}}_1 + n_2\overline{\mathbf{y}}_2}{N} \text{ and } \mathrm{Var}(\overline{\mathbf{y}}_\cdot) = \frac{1}{N}\boldsymbol{\Sigma}.$$

In approach 2, it is not assumed that the profiles are parallel. An average profile can still be obtained as a weighted average of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In this case, however, the choice of w does influence the shape of the average profile. For example, by giving more weight to $\boldsymbol{\mu}_1$, the shape of the average profile will look more like the shape of $\boldsymbol{\mu}_1$ and less like the shape of $\boldsymbol{\mu}_2$. It is conventional, in this case, to define the average profile as follows:

$$\boldsymbol{\mu}_\cdot = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

The corresponding estimator and its variance are

$$\overline{\mathbf{y}}_\cdot = \frac{1}{2}(\overline{\mathbf{y}}_1 + \overline{\mathbf{y}}_2) \text{ and } \mathrm{Var}(\overline{\mathbf{y}}_\cdot) = \frac{N}{4n_1 n_2}\boldsymbol{\Sigma}.$$

Theorem 7.15 (Approach 1 to Shape of Average Profile) *Confidence intervals and hypothesis tests concerning the shape of the average profile can be based on the following pivotal quantity:*

$$\left(\frac{N(N-d)}{(N-2)(d-1)}\right)(\overline{\mathbf{y}}_\cdot - \boldsymbol{\mu})'\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}'(\overline{\mathbf{y}}_\cdot - \boldsymbol{\mu}) \sim F_{d-1,N-d,0},$$

where $\overline{\mathbf{y}}_\cdot = (n_1\overline{\mathbf{y}}_1 + n_2\overline{\mathbf{y}}_2)/N$, $\boldsymbol{\mu} = (n_1\boldsymbol{\mu}_1 + n_2\boldsymbol{\mu}_2)/N$, *and* $\mathbf{S}$ *is given in Theorem 7.13. In particular,* $\mathrm{H}_0\colon \mathbf{M}'\boldsymbol{\mu}_\cdot = \boldsymbol{\delta}$ *is rejected for large* T^2, *where*

$$T^2 = N(\mathbf{M}'\overline{\mathbf{y}}_\cdot - \boldsymbol{\delta})'(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}(\mathbf{M}'\overline{\mathbf{y}} - \boldsymbol{\delta}).$$

The distribution of T^2 *is*

$$\left(\frac{(N-d)}{(N-2)(d-1)}\right)T^2 \sim F_{d-1,N-d,\lambda}, \text{ where } \lambda = N\frac{(<'\boldsymbol{\mu} - \boldsymbol{\delta})'(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}(\mathbf{M}'\boldsymbol{\mu} - \boldsymbol{\delta})}{2}.$$

Simultaneous confidence intervals on linear functions $\boldsymbol{\ell}'\boldsymbol{\mu}_\cdot$ *can be obtained from*

$$\Pr[\boldsymbol{\ell}'\overline{\mathbf{y}}_\cdot - k \leq \boldsymbol{\ell}'\boldsymbol{\mu}_\cdot \leq \boldsymbol{\ell}'\overline{\mathbf{y}}_\cdot + k \;\forall\; \boldsymbol{\ell} \in \mathcal{R}(\mathbf{M})] = 1 - \alpha,$$

where $k = \sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}\, F^*}$, *and*

$$F^* = \left(\frac{(d-1)(N-2)}{N(N-d)}\right) F^{1-\alpha}_{d-1,N-d}.$$

Proof: Use Theorem 7.4.

□

Theorem 7.16 (Approach 2 to Shape of Average Profile) *Confidence intervals and hypothesis tests concerning the shape of the average profile can be based on the following pivotal quantity:*

$$\left(\frac{4n_1n_2(N-d)}{N(N-2)(d-1)}\right)(\overline{\mathbf{y}}_{\cdot}-\boldsymbol{\mu})'\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}\mathbf{M}'(\overline{\mathbf{y}}_{\cdot}-\boldsymbol{\mu})\sim F_{d-1,N-d,0},$$

where $\overline{\mathbf{y}}_{\cdot}=\frac{1}{2}(\overline{\mathbf{y}}_1+\overline{\mathbf{y}}_2))$, $\boldsymbol{\mu}=\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$, *and* $\mathbf{S}$ *is given in Theorem 7.13. In particular,* $\mathrm{H}_0\colon \mathbf{M}'\boldsymbol{\mu}_{\cdot}=\boldsymbol{\delta}$ *is rejected for large* T^2, *where*

$$T^2=\left(\frac{4n_1n_2}{N}\right)(\mathbf{M}'\overline{\mathbf{y}}_{\cdot}-\boldsymbol{\delta})'\mathbf{M}(\mathbf{M}'\mathbf{S}\mathbf{M})^{-1}(\mathbf{M}'\overline{\mathbf{y}}_{\cdot}-\boldsymbol{\delta}).$$

The distribution of T^2 *is*

$$\left(\frac{N-d}{(N-2)(d-1)}\right)T^2\sim F_{d-1,N-d,\lambda},\ \text{where}\ \lambda=\left(\frac{4n_1n_2}{N}\right)\frac{(\mathbf{M}'\boldsymbol{\mu}_{\cdot}-\boldsymbol{\delta})'\mathbf{M}(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}(\mathbf{M}'\boldsymbol{\mu}_{\cdot}-\boldsymbol{\delta})}{2}.$$

Simultaneous confidence intervals on linear functions $\boldsymbol{\ell}'\boldsymbol{\mu}_{\cdot}$ *can be obtained from*

$$\Pr[\boldsymbol{\ell}'\overline{\mathbf{y}}_{\cdot}-k\le\boldsymbol{\ell}'\boldsymbol{\mu}_{\cdot}\le\boldsymbol{\ell}'\overline{\mathbf{y}}_{\cdot}+k\ \forall\ \boldsymbol{\ell}\in\mathcal{R}(\mathbf{M})]=1-\alpha,$$

where $k=\sqrt{\boldsymbol{\ell}'\mathbf{S}\boldsymbol{\ell}\,F^*}$, *and*

$$F^*=\left(\frac{N(d-1)(N-2)}{4n_1n_2(N-d)}\right)F^{1-\alpha}_{d-1,N-d}.$$

□

7.4 SUMMARY OF HOTELLING'S T^2 AND SAS CODE

The model underlying one and two-sample Hotelling T^2 tests, can be written as

$$\mathbf{Y}=\mathbf{X}\mathbf{B}+\mathbf{U},$$

where $\mathrm{vec}(\mathbf{U})\sim\mathrm{N}(\mathbf{0},\boldsymbol{\Sigma}\otimes\mathbf{I}_n)$; $\mathbf{Y}$ is $n\times d$; $\mathbf{X}$ is $n\times p$ with rank-r; and $\mathbf{B}$ is $p\times d$. In all cases, the hypotheses can be written as $\mathrm{H}_0\colon \mathbf{L}\mathbf{B}\mathbf{M}=\boldsymbol{\Delta}_0$ against $\mathrm{H}_a\colon \mathbf{L}\mathbf{B}\mathbf{M}\neq\boldsymbol{\Delta}_0$, where $\mathbf{L}\mathbf{B}$ is an estimable function; $\boldsymbol{\Delta}_0$ is a known matrix (usually equal to zero); $\mathbf{L}$ is $f\times p$ with rank f; and $\mathbf{M}$ is $d\times q$ with rank q. Note that $\mathbf{M}$ could be equal to $\mathbf{I}_d$.

In general, the $\mathbf{H}$ and $\mathbf{E}$ matrices are

$$\mathbf{H}=(\mathbf{L}\widetilde{\mathbf{B}}\mathbf{M}-\boldsymbol{\Delta}_0)'\left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\right]^{-1}(\mathbf{L}\widetilde{\mathbf{B}}\mathbf{M}-\boldsymbol{\Delta}_0)\ \text{and}\ \mathbf{E}=\mathbf{M}'\mathbf{Y}(\mathbf{I}_n-\mathbf{H}_x)\mathbf{Y}\mathbf{M},$$

where $\widetilde{\mathbf{B}}=(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ and $\mathbf{H}_x=\mathrm{ppo}(\mathbf{X})$. The matrices $\mathbf{H}$ and $\mathbf{E}$ are independently distributed as

$$\mathbf{H}\sim\mathrm{W}_q(f,\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M},\boldsymbol{\Lambda})\ \text{and}\ \mathbf{E}\sim\mathrm{W}_q(n-r,\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}),\ \text{where}$$

$$\boldsymbol{\Lambda}=(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}(\mathbf{L}\mathbf{B}\mathbf{M}-\boldsymbol{\Delta}_0)'\left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\right]^{-1}(\mathbf{L}\mathbf{B}\mathbf{M}-\boldsymbol{\Delta}_0).$$

In the special case of one and two-sample Hotelling's T^2, the matrix $\mathbf{L}$ is $1\times p$ so $\mathbf{H}\mathbf{E}^{-1}$ has only one non-zero eigenvalue. In this case,

$$T^2=(n-r)\,\mathrm{trace}\left(\mathbf{H}\mathbf{E}^{-1}\right)\ \text{and}\ \left(\frac{n-r-q+1}{(n-r)q}\right)T^2\sim F_{q,n-r-q+1,\lambda},\ \text{where}$$

$$\lambda=\frac{(\mathbf{L}\mathbf{B}\mathbf{M}-\boldsymbol{\Delta}_0)(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M})^{-1}(\mathbf{L}\mathbf{B}\mathbf{M}-\boldsymbol{\Delta}_0)'}{\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'}.$$

Tables for the usual multivariate test statistics are entered with three parameters: s, m, and n. Caution, n is not sample size; it is defined below. If $\mathbf{E}$ and $\mathbf{H}$ have independent Wishart distributions, $\mathbf{E}\sim\mathrm{W}_q(\nu_E,\boldsymbol{\Sigma})$ and $\mathbf{H}\sim\mathrm{W}_q(\nu_H,\boldsymbol{\Sigma})$, then the tables are entered with

$$s=\min(\nu_H,q),\quad m=\frac{1}{2}(|\nu_H-q|-1),\ \text{and}\ n=\frac{1}{2}(\nu_E-q-1).$$

7.4.1 One Sample Hotelling's T^2

1. Model:
$$\text{vec}(\mathbf{Y}) \sim \text{N}\left[\text{vec}(\mathbf{1}_n\boldsymbol{\mu}'), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)\right], \text{ where } \mathbf{Y} \text{ is } n \times d.$$

2. Conventional Hypotheses

   (a) Test that mean vector is equal to an a priori specified vector: $\text{H}_0\colon \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $\text{H}_a\colon \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

   (b) In repeated measures, test hypotheses about the shape of the the profile of means (i.e., changes over time): $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$ versus $\text{H}_a\colon \mathbf{M}'\boldsymbol{\mu} \neq \mathbf{0}$, where $\mathbf{M}$ is $d \times q$ with rank q and $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$.

   (c) In repeated measures, test hypotheses about the location (i.e., mean) of the profile: $\text{H}_0\colon (1/d)\mathbf{1}_d'\boldsymbol{\mu} = \boldsymbol{\theta}_0$ versus $\text{H}_a\colon (1/d)\mathbf{1}_d'\boldsymbol{\mu} \neq \boldsymbol{\theta}_0$,

3. Minimal SAS Commands for Omnibus Test

   (a) For $\text{H}_0\colon \boldsymbol{\mu} = \mathbf{0}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
proc glm;
  model y1 y2 ... yd = /nouni;
  manova H=intercept/summary;
```

   Compare T^2 to $d(n-1)F^{1-\alpha}_{d,n-d}/(n-d)$.

   (b) Arbitrary Linear Functions: If a test of $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{M}$ is $p \times q$ with rank q, then the $\mathbf{M}$ matrix must be specified. Note $q < d-1$ is allowed and $\mathbf{M}'\mathbf{1}_d$ need not equal $\mathbf{0}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
proc glm;
  model y1 y2 ... yd = /nouni;
  manova H=intercept M = (m11 m21 m31 ... md1,
                          m12 m22 m32 ... md2,
                          ..................,
                          m1q m2q m3q ... mdq)/summary;
```

   Compare T^2 to $\frac{q(n-1)}{n-q}F^{1-\alpha}_{q,n-q}$. Note: it actually is $\mathbf{M}'$ rather than $\mathbf{M}$ that is specified in the manova statement.

   (c) Profile (repeated measures): for $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{M}$ is $d \times (d-1)$ with rank $d-1$ and $\mathbf{M}$ satisfies $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
proc glm;
  model y1 y2 ... yd = /nouni;
  repeated Time d/printm summary;
```

   Compare T^2 to $\frac{(d-1)(n-1)}{n-d+1}F^{1-\alpha}_{d-1,n-d+1}$.

4. Minimal SAS Commands for Follow-up Tests

   (a) Follow-up to $\text{H}_0\colon \boldsymbol{\mu} = \mathbf{0}$. For example, test $\text{H}_0\colon \mu_1 - \mu_2 = 0$ and $\text{H}_0\colon \mu_1 + \mu_2 - \mu_3 = 0$ after rejecting $\text{H}_0\colon \boldsymbol{\mu} = \mathbf{0}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
proc glm;
```

```
  model y1 y2 ... yd = ;
  manova H=intercept m= y1-y2 /summary;
  manova H=intercept m= y1+y2-y3 /summary;
```

Compare T^2 to $\frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
  psi1 = y1-y2;
  psi2 = y1+y2-y3;
proc glm;
  model psi1 psi2 = ;
  estimate 'name1' intercept 1;
```

Compute CI as $\widehat{\psi} \pm \text{se}(\widehat{\psi})\sqrt{\frac{d(n-1)}{n-d}F^{1-\alpha}_{d,n-d}}$.

(b) Arbitrary Linear Functions: follow-up to $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{M}$ is $d \times q$ with rank q and need not satisfy $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$. For example, test $\text{H}_0\colon \mu_1 - \mu_2 = 0$ and $\text{H}_0\colon \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$ after rejecting $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
proc glm;
  model y1 y2 ... yd = /nouni;
  manova H=intercept M = y1-y2/summary;
  manova H=intercept M = y1+y2-y3-y4/summary;
```

The coefficient vectors in the "$M =$" portion of the manova statement must be in the column space of $\mathbf{M}$. Compare T^2 to $\frac{q(n-1)}{n-q}F^{1-\alpha}_{q,n-q}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
  psi1=y1-y2;
  psi2=y1+y2-y3-y4;
proc glm;
  model psi1 psi2 = ;
  estimate 'name1' intercept 1;
```

The coefficient vector for each ψ_j must be in the column space of $\mathbf{M}$. Compute CI as $\widehat{\psi} \pm \text{se}(\widehat{\psi})\sqrt{\frac{q(n-1)}{n-q}F^{1-\alpha}_{q,n-q}}$.

(c) Profile (repeated measures): follow-up to $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$, where $\mathbf{M}$ is $d \times (d-1)$ with rank $p-1$ and satisfies $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$. For example, test $\text{H}_0\colon \mu_1 - \mu_2 = 0$ and $\text{H}_0\colon \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$ after rejecting $\text{H}_0\colon \mathbf{M}'\boldsymbol{\mu} = \mathbf{0}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
proc glm;
  model y1 y2 ... yd = /nouni;
  manova H=intercept M = y1-y2/summary;
  manova H=intercept M = y1+y2-y3-y4/summary;
```

The coefficient vectors in the "$M =$" portion of the manova statement must be contrast coefficient vectors. Compare T^2 to $\frac{(d-1)(n-1)}{n-d+1}F^{1-\alpha}_{d-1,n-d+1}$.

```
data;
  infile datafile;
  input y1 y2 ... yd;
  psi1=y1-y2;
  psi2=y1+y2-y3-y4;
proc glm;
  model psi1 psi2 = ;
  estimate 'name1' intercept 1;
```

The coefficient vector for each ψ_j must be a contrast coefficient vector. Compute CI as $\widehat{\psi} \pm \text{se}(\widehat{\psi})\sqrt{\frac{(d-1)(n-1)}{n-d+1}F^{1-\alpha}_{d-1,n-d+1}}$.

7.4.2 Univariate Profile Analyses

1. Randomized Block (one sample analysis)

   (a) Model:

   $$\text{vec}(\mathbf{Y}) \sim \text{N}[\text{vec}(\mathbf{1}_n\boldsymbol{\mu}'), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)],$$

   where $\mathbf{Y}$ is $n \times d$ and $\boldsymbol{\Sigma}$ satisfies $\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}_{d-1}$ and $\mathbf{M}$ is a $d \times (d-1)$ orthonormal matrix of contrast coefficients. One mixed model that satisfies the sphericity condition is the following:

   $$\mathbf{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{\pi}\mathbf{1}_d' + \mathbf{U},$$

   where $\boldsymbol{\pi} \sim \text{N}(\mathbf{0}, \sigma^2_\pi\mathbf{I}_n)$ and $\text{vec}(\mathbf{U}) \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I}_{nd})$. The scalar form of the mixed model is

   $$y_{ij} = \mu + \pi_i + \alpha_j + \varepsilon_{ij}.$$

   For this model, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p + \sigma^2_\pi\mathbf{1}_d\mathbf{1}_d'$. This structure (equal variances and equal covariances) is known as compound symmetry. The randomized block analysis is valid if and only if $\boldsymbol{\Sigma}$ satisfies $\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}_{d-1}$, where $\mathbf{M}$ is an $d \times (d-1)$ semi-orthogonal matrix of contrast coefficients. That is, $\mathbf{M}'\mathbf{M} = \mathbf{I}_{d-1}$ and $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$.

   (b) ANOVA Table:

Source	*df*
Intercept	1
Factor A (repeated measure)	$d-1$
Subjects	$n-1$
Residual	$(d-1)(n-1)$
Total	nd

   (c) Minimal SAS Commands.

```
data;
   infile datafile;
   input subj Fac_A y;
proc GLM;
   class subj Fac_A;
   model y = subj Fac_A;
   random subj / test;
```

2. Split-Plot Factorial (multi-sample analysis)

   (a) Model:

   $$\text{vec}(\mathbf{Y}) \sim \text{N}[\text{vec}(\mathbf{XB}), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)],$$

   where $\mathbf{Y}$ is $N \times d$, $\boldsymbol{\Sigma}$ satisfies $\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} = \sigma^2\mathbf{I}_{d-1}$, $\mathbf{M}$ is a $d \times (d-1)$ orthonormal matrix of contrast coefficients,

   $$\mathbf{X} = \bigoplus_{j=1}^{a} \mathbf{1}_{n_j}, \quad \mathbf{B} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_a \end{pmatrix},$$

and $N = \sum_{j=1}^{a} n_j$. One mixed model that satisfies the sphericity condition is the following:

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\pi}\mathbf{1}_d' + \mathbf{U},$$

where $\boldsymbol{\pi} \sim \mathrm{N}(\mathbf{0}, \sigma_\pi^2 \mathbf{I}_N)$ and $\mathrm{vec}(\mathbf{U}) \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{Nd})$. The scalar form of the mixed model is

$$y_{ijk} = \mu + \pi_i + \alpha_j + \tau_k + \gamma_{jk} + \varepsilon_{ijk},$$

where α_j for $j = 1, \ldots, a$ are the effects of the between-subjects factor and τ_k for $k = 1, \ldots, d$ are the effects of the within-subjects factor. For this model, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_d + \sigma_\pi^2 \mathbf{1}_d \mathbf{1}_d'$. This structure (equal variances and equal covariances) is known as compound symmetry. The split-plot analysis is valid if and only if $\boldsymbol{\Sigma}$ satisfies $\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} = \sigma^2 \mathbf{I}_{d-1}$, where $\mathbf{M}$ is a $d \times (d-1)$ semi-orthogonal matrix of contrast coefficients. That is, $\mathbf{M}'\mathbf{M} = \mathbf{I}_{d-1}$ and $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$.

(b) ANOVA Table:

Source	*df*
Intercept	1
Between Subjects	$N-1$
Factor A	$a-1$
Subjects within A	$N-a$
Within Subjects	$N(d-1)$
Factor B (repeated measure)	$d-1$
AB Interaction	$(a-1)(d-1)$
Residual	$(d-1)(N-a)$
Total	Nd

(c) Minimal SAS Commands.

```
data;
   infile datafile;
   input subj Fac_A Fac_B y;
proc GLM;
   class subj Fac_A Fac_B;
   model y = subj(Fac_A) Fac_A|Fac_B;
   random subj(Fac_A) / test;
```

7.4.3 Two Sample Hotelling's T^2

1. Model:

$$\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}\left[\mathrm{vec}(\mathbf{XB}), (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)\right], \text{ where } \mathbf{Y} \text{ is } n \times d,$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_N & \bigoplus_{j=1}^{2} \mathbf{1}_{n_j} \end{pmatrix}, \text{ and } \mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_0' \\ \boldsymbol{\tau}_1' \\ \boldsymbol{\tau}_2' \end{pmatrix}.$$

Note: $\boldsymbol{\mu}_i = \boldsymbol{\beta}_0 + \boldsymbol{\tau}_i$ for $i = 1, 2$.

2. Hypotheses: $\mathrm{H}_0 \colon \mathbf{LBM} = \boldsymbol{\Delta}_0$ against $\mathrm{H}_a \colon \mathbf{LBM} \neq \boldsymbol{\Delta}_0$

   (a) To average over groups, use $\mathbf{L} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. To compare groups, use $\mathbf{L} = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$.

   (b) To average over repeated measures (i.e., time), use $\mathbf{M} = (1/d)\mathbf{1}_d$. To examine differences among repeated measures, use

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

   (c) Test equality of mean vectors: $\mathrm{H}_0 \colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $\mathrm{H}_a \colon \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ or, equivalently, $\mathrm{H}_0 \colon \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$ versus $\mathrm{H}_a \colon \boldsymbol{\tau}_1 \neq \boldsymbol{\tau}_2$. Use $\mathbf{L} = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$ and $\mathbf{M} = \mathbf{I}_p$.

(d) Profile Hypotheses:

i. Test equal Locations of Profiles (main effect of treatment): $H_0: \mathbf{1}_d'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$ versus $H_a: \mathbf{1}_d'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \neq 0$. Use $\mathbf{L} = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$ and $\mathbf{M} = \mathbf{1}_d$.

ii. Test that average Profile (over two groups) is flat (zero changes over time). This is the main effect for time. Hypotheses are $H_0: \mathbf{M}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 = \mathbf{0}$ versus $H_a: \mathbf{M}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2 \neq \mathbf{0}$, where $\mathbf{L} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$, $\mathbf{M}$ is $d \times (d-1)$ with rank $d-1$, and $\mathbf{M}$ satisfies $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$.

iii. Test that the two profiles are parallel (treatment $\times$ time interaction). Hypotheses are $H_0: \mathbf{M}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ versus $H_a: \mathbf{M}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \neq \mathbf{0}$, where $\mathbf{L} = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}$, $\mathbf{M}$ is $d \times (d-1)$ with rank $d-1$, and $\mathbf{M}$ satisfies $\mathbf{M}'\mathbf{1}_d = \mathbf{0}$.

3. Minimal SAS Commands for Omnibus Test

(a) For $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

```
data;
  infile datafile;
  input treat y1 y2 ... yd;
proc glm;
  class treat;
  model y1 y2 ... yd = treat/nouni;
  manova H=treat/summary;
```

Compare T^2 to $d(N-2)F^{1-\alpha}_{d,N-d-1}/(N-d-1)$.

(b) Profile Analysis (all three tests: equal location, average profile is flat, & parallel profiles)

```
data;
  infile datafile;
  input treat y1 y2 ... yd;
proc glm;
  class treat;
  model y1 y2 ... yd = treat /nouni;
  repeated Time d /summary;
```

Compare T_1^2 (for equal levels) to $F^{1-\alpha}_{1,N-2}$. Compare T_2^2 (zero slope for average profile) to $(d-1)(N-2)F^{1-\alpha}_{d-1,N-d}/(N-d)$. Compare T_3^2 (parallel profiles) to $(d-1)(N-2)F^{1-\alpha}_{d-1,N-d}/(N-d)$.

4. Minimal SAS Commands for Follow-up Tests

(a) Follow-up to $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

```
data;
  infile datafile;
  input treat y1 y2 ... yd;
proc glm;
  class treat;
  model y1 y2 ... yd = treat/nouni;
  manova H=treat m=(m1 m2 ... md)/summary;
```

Compare T^2 to $d(N-2)F^{1-\alpha}_{d,N-d-1}/(N-d-1)$.

```
data;
  infile datafile;
  input treat y1 y2 ... yd;
  psi1 = y1-y2;
  psi2 = y1 + y2 -y3;
proc glm;
  class treat;
  model psi1 psi2 = treat;
  estimate 'name' treat 1 -1;
```

Compute CI as $\widehat{\psi} \pm \text{se}(\widehat{\psi})\sqrt{d(N-2)F^{1-\alpha}_{d,N-d-1}/(N-d-1)}$.

(b) Profile (for zero slope of average profile and parallel profiles):

```
data;
  infile datafile;
  input treat y1 y2 ... yd;
proc glm;
  class treat;
  model y1 y2 ... yd = treat /intercept nouni;
  manova H= _all_ M = y1-y2/summary;
  manova H= _all_ M = y1+y2-y3-y4/summary;
```

The coefficient vector $\mathbf{m}$ in the $M =$ component of the manova statement must be a contrast coefficient vector. Compare T_2^2 (zero slopes for average profile) to $(d-1)(N-2)F^{1-\alpha}_{d-1,N-d}/(N-d)$. Compare T_3^2 (parallel profiles) to $(d-1)(N-2)F^{1-\alpha}_{d-1,N-d}/(N-d)$.

```
data;
  infile datafile;
  input treat y1 y2 ... yd;
  psi0 = (y1 + y2 + y3 + ... + yd)/d;
  psi1 = y1-y2;
  psi2 = y1+y2-y3-y4;
proc glm;
  class treat;
  model psi0 psi1 psi2 = treat /nouni;
  estimate 'name1' intercept 1;
  estimate 'name2' treat 1 -1;
```

Compute CI as $\widehat{\psi} \pm \text{se}(\widehat{\psi})\sqrt{(d-1)(N-2)F^{1-\alpha}_{d-1,N-d}/(N-d)}$.

Chapter 8

MULTIVARIATE LINEAR MODELS

8.1 MODEL DESCRIPTION

Consider the standard multivariate linear model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where $\mathbf{Y}$ is $n \times d$:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix},$$

$\mathbf{X}$ is a known $n \times p$ matrix with rank r, $\mathbf{B}$ is an unknown $p \times d$ matrix of parameters, and $\mathbf{U}$ is an $n \times d$ random matrix with distribution $\text{vec}(\mathbf{U}) \sim [\mathbf{0}, (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$. It can be deduced that the rows of $\mathbf{Y}$ are uncorrelated with one another and $\mathbf{y}_i \sim (\mathbf{B}'\mathbf{x}_i, \mathbf{\Sigma})$, where $\mathbf{x}_i$ is the i^{th} row of $\mathbf{X}$. The vector equivalent of the model is

$$\text{vec}(\mathbf{Y}) = \mathbf{y} = (\mathbf{I}_d \otimes \mathbf{X})\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$ and $\boldsymbol{\varepsilon} = \text{vec}(\mathbf{U})$.

8.2 ESTIMABILITY & BLUES

From prior work, we know that the GLSE (MLE under normality) of $\mathbf{B}$ is

$$\widetilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

and that

$$\mathbf{S} = \frac{\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}}{n - r}$$

is unbiased for $\mathbf{\Sigma}$ where $\mathbf{H}_x = \text{ppo}(\mathbf{X})$. In fact, under normality, $\mathbf{S}$ is the UMVUE of $\mathbf{\Sigma}$.

To estimate a linear function of $\mathbf{B}$, say $\mathbf{L}'\mathbf{BM}$, we would like to be able to use the natural estimator, $\mathbf{L}'\widetilde{\mathbf{B}}\mathbf{M}$. It turns out that this is a sensible thing to do, provided that $\mathbf{L}'\mathbf{BM}$ is estimable. The function $\mathbf{\Psi} = \mathbf{L}'\mathbf{BM}$ is estimable if there exists a linear unbiased estimator of $\mathbf{\Psi}$: $\widehat{\mathbf{\Psi}} = \mathbf{FYG} + \mathbf{K}$, and $\text{E}(\widehat{\mathbf{\Psi}}) = \mathbf{\Psi}$ for all $\mathbf{B}$.

Theorem 8.1 *Estimability: The function $\mathbf{\Psi} = \mathbf{L}'\mathbf{BM}$ is estimable if and only if $\mathbf{L} \in \mathcal{R}(\mathbf{X}')$. That is, if the columns of $\mathbf{L}$ are $\mathbf{l}_1, \ldots, \mathbf{l}_s$, then $\mathbf{\Psi}$ is estimable if and only if $\mathbf{l}_i \in \mathcal{R}(\mathbf{X}')$ for all i.*
Proof HW or in class.

□

8.2.1 BLUE

Definition: Best Linear Unbiased Estimator (BLUE)—Scalar. Consider the univariate linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ with rank-r and $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Omega})$. Suppose that $\mathbf{l}'\boldsymbol{\beta}$ is an estimable function. That is $\mathbf{l}$ is a $p \times 1$ vector and $\mathbf{l} \in \mathcal{R}(\mathbf{X}')$. Let $\widehat{\psi}_{\mathbf{l}}$ be a linear unbiased estimator of $\mathbf{l}'\boldsymbol{\beta}$. That is,

1. $\widehat{\psi}_{\mathbf{l}} = \mathbf{a}'\mathbf{y} + k$ for some $\mathbf{a}$: $n \times 1$ and some scalar k, and
2. $\mathrm{E}(\widehat{\psi}_{\mathbf{l}}) = \mathbf{l}'\boldsymbol{\beta}$.

A linear unbiased estimator is said to be the best linear unbiased estimator (BLUE) if it has the minimum variance over all linear unbiased estimators.

Definition: Best Linear Unbiased Estimator (BLUE)—Vector. Consider the univariate linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ with rank-r and $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Omega})$. Suppose that $\mathbf{L}'\boldsymbol{\beta}$ is a vector of estimable functions. That is $\mathbf{L}$ is a $p \times q$ matrix and $\mathbf{L} \in \mathcal{R}(\mathbf{X}')$. Let $\widehat{\boldsymbol{\psi}}_{\mathbf{L}}$ be a linear unbiased estimator of $\mathbf{L}'\boldsymbol{\beta}$. That is,

1. $\widehat{\boldsymbol{\psi}}_{\mathbf{L}} = \mathbf{L}'\mathbf{y} + \mathbf{k}$ for some $\mathbf{L}$: $n \times q$ and some vector $\mathbf{k}$: $q \times 1$, and
2. $\mathrm{E}(\widehat{\boldsymbol{\psi}}_{\mathbf{L}}) = \mathbf{L}'\boldsymbol{\beta}$.

A vector of linear unbiased estimators is said to be BLUE if each entry in the vector is BLUE. Denote the $q \times q$ covariance matrix of $\widehat{\boldsymbol{\psi}}_{\mathbf{L}}$ by $\mathbf{V}$. Let $\widehat{\boldsymbol{\psi}}^*_{\mathbf{L}}$ be another linear unbiased estimator of $\mathbf{L}'\boldsymbol{\beta}$. Denote the $q \times q$ covariance matrix of $\widehat{\boldsymbol{\psi}}^*_{\mathbf{L}}$ by $\mathbf{V}^*$. Then, $\widehat{\boldsymbol{\psi}}_{\mathbf{L}}$ is BLUE iff $\mathrm{tr}(\mathbf{V}) \leq \mathrm{tr}(\mathbf{V}^*)$ for all $\mathbf{V}^*$.

Theorem 8.2 *Gauss-Markov Consider the univariate linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ with rank-r, $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\boldsymbol{\Omega})$, and $\boldsymbol{\Omega} > 0$. Let $\mathbf{L}$: $p \times q$ be a matrix of constants satisfying $\mathbf{L} \in \mathcal{R}(\mathbf{X}')$. Then $\mathbf{L}'\widetilde{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator (BLUE) of $\mathbf{L}'\boldsymbol{\beta}$ where $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$. Note: this result does not depend on normality.*

Proof: Stat 505-506 Homework

□

1. Corollary 1: $\mathrm{Var}(\mathbf{L}'\widetilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{L}'(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-}\mathbf{L}$.
2. Corollary 2: If $\boldsymbol{\Omega} = \mathbf{I}_n$, then the BLUE of $\mathbf{L}'\boldsymbol{\beta}$ is $\mathbf{L}'\widetilde{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$, and $\mathrm{Var}(\mathbf{L}'\widetilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}$.
3. Corollary 3: Consider the usual multivariate setup: $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}$, where $\mathrm{vec}(\mathbf{U}) \sim [\mathbf{0}, (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)]$. If $\mathbf{L} \in \mathcal{R}(\mathbf{X}')$, then the BLUE of $\mathbf{L}'\mathbf{B}\mathbf{M}$ is $\mathbf{L}'\widetilde{\mathbf{B}}\mathbf{M}$, $\widetilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$, and $\mathrm{disp}(\mathbf{L}'\widetilde{\mathbf{B}}\mathbf{M}) = [\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} \otimes \mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}]$.
4. Corollary 4: Consider the setup in Corollary 2.3. If $\mathbf{X} = \mathbf{1}_n$ and $\mathbf{B} = \boldsymbol{\mu}'$, then the BLUE of $\boldsymbol{\mu}$ is $\overline{\mathbf{y}} = n^{-1}\mathbf{Y}'\mathbf{1}_n$ and $\mathrm{var}(\overline{\mathbf{y}}) = n^{-1}\boldsymbol{\Sigma}$.

We now turn to the development of inference procedures on linear functions of $\mathbf{B}$. We will consider hypotheses of the form H_0: $\mathbf{L}'\mathbf{B}\mathbf{M} = \boldsymbol{\Delta}$. Often $\boldsymbol{\Delta}$ will be $\mathbf{0}$ and $\mathbf{M}$ will be $\mathbf{I}$. It will be assumed that the rows of $\mathbf{Y}$ independently follow a multivariate normal distribution.

8.3 ESTIMATING B AND Σ UNDER CONSTRAINTS

To construct the likelihood ratio test of H_0: $\mathbf{L}'\mathbf{B}\mathbf{M} = \boldsymbol{\Delta}$, it will be necessary to maximize the likelihood function subject to the constraint $\mathbf{L}'\mathbf{B}\mathbf{M} = \boldsymbol{\Delta}$.

8.3.1 Case I: $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$, where $\mathbf{M}$ is Non-Singular

Theorem 8.3 *Let $\mathbf{C}$ be a $p \times q$ matrix with rank q and suppose that $\mathbf{C}'\mathbf{B}$ is estimable. Let $\mathbf{M}$ be a known $d \times d$ matrix with rank d. The generalized least squares estimator of $\mathbf{B}$ subject to the restriction that $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ is given by*

$$\widetilde{\mathbf{B}}_0 = \widetilde{\mathbf{B}}_a - (\mathbf{X}'\mathbf{X})^-\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M} - \mathbf{\Delta})\mathbf{M}^{-1},$$

where $\widetilde{\mathbf{B}}_a = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$ is the ordinary unrestricted estimator of $\mathbf{B}$ (under H_a).

Outline of Proof: To compute the maximizer with respect to $\mathbf{B}$, solve

$$\frac{\partial Q}{\partial \operatorname{vec}(\mathbf{B})} = \mathbf{0} \text{ and } \frac{\partial Q}{\partial \boldsymbol{\lambda}} = \mathbf{0}$$

where

$$Q = \operatorname{tr}\left[(\mathbf{Y} - \mathbf{X}\mathbf{B})'(\mathbf{Y} - \mathbf{X}\mathbf{B})\mathbf{\Sigma}^{-1}\right] - 2\boldsymbol{\lambda}'\left[(\mathbf{M}' \otimes \mathbf{C}')\boldsymbol{\beta} - \boldsymbol{\delta}\right],$$

$\boldsymbol{\beta} = \operatorname{vec}(\mathbf{B})$; $\boldsymbol{\lambda} = \operatorname{vec}(\mathbf{\Lambda})$; and $\boldsymbol{\delta} = \operatorname{vec}(\mathbf{\Delta})$.

□

Note 1: If the rank of $\mathbf{X}$ is p, then $(\mathbf{X}'\mathbf{X})^-$ can be replaced by $(\mathbf{X}'\mathbf{X})^{-1}$ and $\widetilde{\mathbf{B}}_0$ will be the BLUE of $\mathbf{B}$.
Note 2: Under normality and a true H_0, $\widetilde{\mathbf{B}}_0$ is the MLE of $\mathbf{B}$.

Theorem 8.4 *Assuming normality of $\mathbf{Y}$ and non-singularity of $\mathbf{M}$, the MLE of $\mathbf{\Sigma}$ under H_0: $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ is*

$$\widetilde{\mathbf{\Sigma}}_0 = \widetilde{\mathbf{\Sigma}}_a + \mathbf{M}'^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M} - \mathbf{\Delta})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M} - \mathbf{\Delta})\mathbf{M}^{-1}/n,$$

where $\widehat{\mathbf{\Sigma}}_a = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}/n$.

Proof: When $\mathbf{B}$ is estimated under the constraint $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$, the residual sum of squares and cross products matrix is

$$\begin{aligned}
\mathbf{T} &= (\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_0)'(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_0) \\
&= \left[(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_a) + \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})\mathbf{M}^{-1}\right]' \\
&\quad \left[(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_a) + \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})\mathbf{M}^{-1}\right] \\
&= \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y} + \mathbf{M}'^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})\mathbf{M}^{-1},
\end{aligned}$$

where $\mathbf{H}_x = \operatorname{ppo}(\mathbf{X})$. From prior results, the MLE of $\mathbf{\Sigma}$ is $\mathbf{\Sigma}_0 = \mathbf{T}/n$.

□

Note that

$$\begin{aligned}
\mathrm{E}(\mathbf{T}) &= (n - r + q)\mathbf{\Sigma} + \mathbf{M}'^{-1}(\mathbf{C}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})\mathbf{M}^{-1} \\
&= (n - r + q)\mathbf{\Sigma} \text{ if } \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}.
\end{aligned}$$

Accordingly, if $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ is true, then $\mathbf{A}/(n - r + q)$ is an unbiased estimator of $\mathbf{\Sigma}$. This result does not require normality.

8.3.2 Case II: $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$, where $\mathbf{M}$ is not Square

Theorem 8.5 *Let $\mathbf{C}$ be a $p \times q$ matrix with rank q and suppose that $\mathbf{C}'\mathbf{B}$ is estimable. Let $\mathbf{M}$ be a known $d \times k$ matrix with rank k. Further, suppose that $\mathbf{\Sigma}$ is known. Then, the generalized least squares estimator of $\mathbf{B}$ (MLE under normality), subject to the restriction $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$, is given by*

$$\widetilde{\mathbf{B}}_0 = \widetilde{\mathbf{B}}_a - (\mathbf{X}'\mathbf{X})^-\mathbf{C}\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\mathbf{\Sigma}\mathbf{M})^{-1}\mathbf{M}'\mathbf{\Sigma},$$

where $\widehat{\mathbf{\Delta}} = \mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M}$ and $\widetilde{\mathbf{B}}_a = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$.

Outline of Proof: To compute the constrained GLS estimator, solve

$$\frac{\partial Q}{\partial \operatorname{vec}(\mathbf{B})} = \mathbf{0} \text{ and } \frac{\partial Q}{\partial \boldsymbol{\lambda}} = \mathbf{0}$$

where

$$Q = \operatorname{tr}\left[(\mathbf{Y}-\mathbf{X}\mathbf{B})'(\mathbf{Y}-\mathbf{X}\mathbf{B})\boldsymbol{\Sigma}^{-1}\right] - 2\boldsymbol{\lambda}'\left[(\mathbf{M}'\otimes\mathbf{C}')\boldsymbol{\beta}-\boldsymbol{\delta}\right],$$

$\boldsymbol{\beta} = \operatorname{vec}(\mathbf{B})$*;* $\boldsymbol{\lambda} = \operatorname{vec}(\boldsymbol{\Lambda})$*; and* $\boldsymbol{\delta} = \operatorname{vec}(\boldsymbol{\Delta})$*.*

□

Theorem 8.6 (Constrained MLEs) *Let* $\mathbf{C}$ *be a* $p \times q$ *matrix with rank* q *and suppose that* $\mathbf{C}'\mathbf{B}$ *is estimable. Let* $\mathbf{M}$ *be a known* $d \times k$ *matrix with rank* k*. Then, MLEs of* $\mathbf{B}$ *and* $\boldsymbol{\Sigma}$*, subject to the restriction* $\mathbf{C}'\mathbf{B}\mathbf{M} = \boldsymbol{\Delta}$*, are given by*

$$\begin{aligned}
\widetilde{\mathbf{B}}_0 &= \widetilde{\mathbf{B}}_a - (\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})(\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a, \text{ and} \\
\widetilde{\boldsymbol{\Sigma}}_0 &= \widetilde{\boldsymbol{\Sigma}}_a \\
&+ n^{-1}\widetilde{\boldsymbol{\Sigma}}_a\mathbf{M}(\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})(\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a.
\end{aligned}$$

where $\widehat{\boldsymbol{\Delta}} = \mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M}$*,* $\widetilde{\mathbf{B}}_a = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$*;* $\widetilde{\boldsymbol{\Sigma}}_a = n^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}$*; and* $\mathbf{H}_x = \operatorname{ppo}(\mathbf{X})$*.*
Outline of proof: Let $\mathbf{G} = \begin{pmatrix}\mathbf{M} & \mathbf{R}\end{pmatrix}$*, where* $\mathbf{R}$ *is a* $d \times (d-k)$ *matrix chosen to satisfy*

$$\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a\mathbf{R} = \mathbf{0} \text{ and } \operatorname{rank}(\mathbf{G}) = d.$$

The matrix $\mathbf{R}$ *can be generated by* $\mathbf{R} = \widetilde{\boldsymbol{\Sigma}}_a^{-\frac{1}{2}}\operatorname{null}(\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a^{\frac{1}{2}})$*, where* $\operatorname{null}(\mathbf{W})$ *generates a basis set of vectors for the null space of* $\mathbf{W}$ *(see the MATLAB null command). Note that*

$$\mathbf{G}^{-1} = \begin{bmatrix}(\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\boldsymbol{\Sigma}}_a \\ (\mathbf{R}'\widetilde{\boldsymbol{\Sigma}}_a\mathbf{R})^{-1}\mathbf{R}'\widetilde{\boldsymbol{\Sigma}}_a\end{bmatrix}.$$

Write the log likelihood function as

$$\begin{aligned}
\ell(\mathbf{B},\boldsymbol{\Sigma}|\mathbf{Y}) &= -\frac{1}{2}\operatorname{tr}\left[(\mathbf{Y}-\mathbf{X}\mathbf{B})'(\mathbf{Y}-\mathbf{X}\mathbf{B})\boldsymbol{\Sigma}^{-1}\right] - \frac{n}{2}\ln|\boldsymbol{\Sigma}| \\
&\quad \text{plus terms that do not depend on } \mathbf{B} \text{ and } \boldsymbol{\Sigma} \\
&= -\frac{1}{2}\operatorname{tr}\left[\mathbf{G}'^{-1}\mathbf{G}'(\mathbf{Y}-\mathbf{X}\mathbf{B})'(\mathbf{Y}-\mathbf{X}\mathbf{B})\mathbf{G}\mathbf{G}^{-1}\boldsymbol{\Sigma}^{-1}\right] - \frac{n}{2}\ln|\mathbf{G}'\boldsymbol{\Sigma}\mathbf{G}| \\
&\quad +n\ln|\mathbf{G}| \\
&= -\frac{1}{2}\operatorname{tr}\left[(\mathbf{Z}-\mathbf{X}\boldsymbol{\Theta})'(\mathbf{Z}-\mathbf{X}\boldsymbol{\Theta})\boldsymbol{\Omega}^{-1}\right] - \frac{n}{2}\ln|\boldsymbol{\Omega}| + n\ln|\mathbf{G}| \\
&= \ell(\boldsymbol{\Theta},\boldsymbol{\Omega}|\mathbf{Z}),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{Z} &= \mathbf{Y}\mathbf{G} = \begin{pmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{pmatrix} = \begin{pmatrix}\mathbf{Y}\mathbf{M} & \mathbf{Y}\mathbf{R}\end{pmatrix} \\
\boldsymbol{\Theta} &= \mathbf{B}\mathbf{G} = \begin{pmatrix}\mathbf{B}\mathbf{M} & \mathbf{B}\mathbf{R}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Theta}_1 & \boldsymbol{\Theta}_2\end{pmatrix} \text{ and} \\
\boldsymbol{\Omega} &= \mathbf{G}'\boldsymbol{\Sigma}\mathbf{G} = \begin{pmatrix}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22}\end{pmatrix} = \begin{bmatrix}\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M} & \mathbf{M}'\boldsymbol{\Sigma}\mathbf{R} \\ \mathbf{R}'\boldsymbol{\Sigma}\mathbf{M} & \mathbf{R}'\boldsymbol{\Sigma}\mathbf{R}\end{bmatrix},
\end{aligned}$$

The MLEs of $\mathbf{B}$ *and* $\mathbf{\Sigma}$ *will be obtained by first maximizing* $\ell(\mathbf{\Theta}, \mathbf{\Omega}|\mathbf{Z})$ *with respect to* $\mathbf{\Theta}$ *and* $\mathbf{\Omega}$ *subject to* $\mathbf{C}'\mathbf{\Theta}\mathbf{G}^{-1}\mathbf{M} = \mathbf{\Delta}$. *Denote the maximizers of* $\ell(\mathbf{\Theta}, \mathbf{\Omega}|\mathbf{Z})$ *by* $\widetilde{\mathbf{\Theta}}$ *and* $\widetilde{\mathbf{\Omega}}$. *Denote the maximizers of* $\ell(\mathbf{B}, \mathbf{\Sigma}|\mathbf{Y})$ *by* $\widetilde{\mathbf{B}}$ *and* $\widetilde{\mathbf{\Sigma}}$. *By the invariance property of MLEs*

$$\widetilde{\mathbf{B}} = \widetilde{\mathbf{\Theta}}\mathbf{G}^{-1} \text{ and } \widetilde{\mathbf{\Sigma}} = \mathbf{G}'^{-1}\widetilde{\mathbf{\Omega}}\mathbf{G}^{-1}.$$

Recall, if

$$\text{vec}(\mathbf{Z}) \sim \text{N}\left[\text{vec}(\mathbf{X}\mathbf{\Theta}), \mathbf{\Omega} \otimes \mathbf{I}_n\right]$$

then

$$\begin{aligned}
\text{vec}(\mathbf{Z}_1) &\sim \text{N}\left[\text{vec}(\mathbf{X}\mathbf{\Theta}_1), \mathbf{\Omega}_{11} \otimes \mathbf{I}_n\right] \text{ and} \\
\text{vec}(\mathbf{Z}_2)|\mathbf{Z}_1 &\sim \text{N}\left[\text{vec}(\mathbf{X}\mathbf{\Theta}_{2\cdot1} + \mathbf{Z}_1\mathbf{\Gamma}), \mathbf{\Omega}_{22\cdot1} \otimes \mathbf{I}_n\right], \text{ where} \\
\mathbf{\Theta}_{2\cdot1} &= \mathbf{\Theta}_2 - \mathbf{\Theta}_1\mathbf{\Gamma}; \\
\mathbf{\Gamma} &= \mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}; \text{ and} \\
\mathbf{\Omega}_{22\cdot1} &= \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}.
\end{aligned}$$

The relationship between $(\mathbf{B}, \mathbf{\Sigma})$ *and* $(\mathbf{\Theta}, \mathbf{\Gamma}, \mathbf{\Omega})$ *is*

$$\begin{aligned}
\mathbf{B} &= \mathbf{\Theta}\mathbf{G}^{-1} = \begin{pmatrix}\mathbf{\Theta}_1 & \mathbf{\Theta}_2\end{pmatrix}\mathbf{G}^{-1} \\
&= \begin{pmatrix}\mathbf{\Theta}_1 & \mathbf{\Theta}_{2\cdot1} + \mathbf{\Theta}_1\mathbf{\Gamma}\end{pmatrix}\mathbf{G}^{-1} \text{ and} \\
\mathbf{\Sigma} &= \mathbf{G}'^{-1}\mathbf{\Omega}\mathbf{G}^{-1} = \mathbf{G}'^{-1}\begin{pmatrix}\mathbf{\Omega}_{11} & \mathbf{\Omega}_{11}\mathbf{\Gamma} \\ \mathbf{\Gamma}'\mathbf{\Omega}_{11} & \mathbf{\Omega}_{22\cdot1} + \mathbf{\Gamma}'\mathbf{\Omega}_{11}\mathbf{\Gamma}\end{pmatrix}\mathbf{G}^{-1}.
\end{aligned}$$

Using the above factorization of the density, the log likelihood can be written as

$$\ell(\mathbf{\Theta}, \mathbf{\Omega}|\mathbf{Z}) = \ell_1(\mathbf{\Theta}_1, \mathbf{\Omega}_{11}|\mathbf{Z}_1) + \ell_{2\cdot1}(\mathbf{\Theta}_{2\cdot1}, \mathbf{\Gamma}, \mathbf{\Omega}_{22\cdot1}|\mathbf{Z}_2, \mathbf{Z}_1),$$

where

$$\begin{aligned}
\ell_1(\mathbf{\Theta}_1, \mathbf{\Omega}_{11}|\mathbf{Z}_1) &= -\frac{1}{2}\,\text{tr}\left[(\mathbf{Z}_1 - \mathbf{X}\mathbf{\Theta}_1)'(\mathbf{Z}_1 - \mathbf{X}\mathbf{\Theta}_1)\mathbf{\Omega}_{11}^{-1}\right] - \frac{n}{2}\ln|\mathbf{\Omega}_{11}| \text{ and} \\
\ell_{2\cdot1}(\mathbf{\Theta}_{2\cdot1}, \mathbf{\Gamma}, \mathbf{\Omega}_{22\cdot1}|\mathbf{Z}_2, \mathbf{Z}_1) &= -\frac{1}{2}\,\text{tr}\left[(\mathbf{Z}_2 - \mathbf{X}\mathbf{\Theta}_{2\cdot1} - \mathbf{Z}_1\mathbf{\Gamma})'(\mathbf{Z}_2 - \mathbf{X}\mathbf{\Theta}_{2\cdot1} - \mathbf{Z}_1\mathbf{\Gamma})\mathbf{\Omega}_{22\cdot1}^{-1}\right] \\
&\quad -\frac{n}{2}\ln|\mathbf{\Omega}_{22\cdot1}|.
\end{aligned}$$

To maximize the log likelihood function, the two components can be maximized separately. First maximize ℓ_1 *with respect to* $\mathbf{\Theta}_1$ *and* $\mathbf{\Omega}_{11}$ *subject to the constraint* $\mathbf{C}'\mathbf{\Theta}_1 = \mathbf{\Delta}$. *Second, maximize* $\ell_{2\cdot1}$ *with respect to* $\mathbf{\Theta}_{2\cdot1}$, $\mathbf{\Gamma}$, *and* $\mathbf{\Omega}_{22\cdot1}$, *subject to no constraints. The results are as follows:*

$$\begin{aligned}
\widetilde{\mathbf{\Theta}}_1 &= \widetilde{\mathbf{B}}_a\mathbf{M} - (\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}), \\
\widetilde{\mathbf{\Omega}}_{11} &= \mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M} + n^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}), \\
\widetilde{\mathbf{\Gamma}} &= (\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R} = \mathbf{0}, \\
\widetilde{\mathbf{\Theta}}_{2\cdot1} &= \widetilde{\mathbf{B}}_a\mathbf{R} - \widetilde{\mathbf{B}}_a\mathbf{M}\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{B}}_a\mathbf{R}, \text{ and} \\
\widetilde{\mathbf{\Omega}}_{22\cdot1} &= n^{-1}\mathbf{Z}_2'(\mathbf{I}_n - \mathbf{H}_{x,z_1})\mathbf{Z}_2 = n^{-1}\mathbf{Z}_2'(\mathbf{I}_n - \mathbf{H}_x - \mathbf{H}_{z_1\cdot x})\mathbf{Z}_2
\end{aligned}$$

$$= \mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R} - \mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R} = \mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R},$$

where $\mathbf{H}_{x,z_1} = \text{ppo}\left[\begin{pmatrix}\mathbf{X} & \mathbf{Z}_1\end{pmatrix}\right]$, *and* $\mathbf{H}_{z_1\cdot x} = \text{ppo}\left[(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Z}_1\right]$. *The proof is completed by using the invariance of MLEs and piecing together* $\widetilde{\mathbf{B}}$ *and* $\widetilde{\mathbf{\Sigma}}$ *from the components* $\widetilde{\mathbf{\Theta}}$, $\widetilde{\mathbf{\Gamma}}$, *and* $\widetilde{\mathbf{\Omega}}$. *The result is*

$$\begin{aligned}
\widetilde{\mathbf{B}}_0 &= \widetilde{\mathbf{\Theta}}\mathbf{G}^{-1} = \widetilde{\mathbf{B}}_a\left[\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a + \mathbf{R}(\mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R})^{-1}\mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\right] \\
&\quad -(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a \\
&= \widetilde{\mathbf{B}}_a - (\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a, \text{ and} \\
\widetilde{\mathbf{\Sigma}}_0 &= \mathbf{G}'^{-1}\widetilde{\mathbf{\Omega}}\mathbf{G}^{-1} \\
&= \mathbf{G}'^{-1}\begin{pmatrix}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M} + n^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R}\end{pmatrix} \\
&= \widetilde{\mathbf{\Sigma}}_a\left[\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a + \mathbf{R}(\mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R})^{-1}\mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\right] \\
&\quad +n^{-1}\widetilde{\mathbf{\Sigma}}_a\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a \\
&= \widetilde{\mathbf{\Sigma}}_a \\
&\quad +n^{-1}\widetilde{\mathbf{\Sigma}}_a\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a,
\end{aligned}$$

where $\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a + \mathbf{R}(\mathbf{R}'\widetilde{\mathbf{\Sigma}}_a\mathbf{R})^{-1}\mathbf{R}'\widetilde{\mathbf{\Sigma}}_a = \mathbf{I}_d$ *has been used. This is a projection operator with full rank and, therefore, must be the identity.*

□

8.4 LIKELIHOOD RATIO TEST OF $\mathrm{H}_0\colon \mathbf{L}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$

8.4.1 Case I: M is $d \times d$ with rank d

All the work for constructing the LR test has been done. All that remains is to plug the estimators under H_0 and H_a into the likelihood function. The result is the following.

Theorem 8.7 *Assuming normality, the LR test of* $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *against* $\mathrm{H}_a\colon \mathbf{C}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta}$ *is to reject* H_0 *for small values of*

$$U = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|},$$

where

$$\mathbf{E} = \mathbf{M}'\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}\mathbf{M},$$

and

$$\mathbf{H} = (\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta}).$$

Proof: Substitute $(\widetilde{\mathbf{B}}, \widehat{\mathbf{\Sigma}}_a)$ *and* $(\widetilde{\mathbf{B}}_0, \widehat{\mathbf{\Sigma}}_0)$ *into the likelihood function and use properties of determinants.*

□

To obtain critical values of the above LR statistic, the following result can be used.

Theorem 8.8 *The matrices* $\mathbf{E}$ *and* $\mathbf{H}$ *are independently distributed as*

$$\mathbf{E} \sim \mathrm{W}_d(n-r, \mathbf{M}'\mathbf{\Sigma}\mathbf{M}, \mathbf{0})$$

and

$$\mathbf{H} \sim \mathrm{W}_d(q, \mathbf{M}'\mathbf{\Sigma}\mathbf{M}, \mathbf{\Lambda}),$$

where

$$\mathbf{\Lambda} = (\mathbf{M}'\mathbf{\Sigma}\mathbf{M})^{-1}(\mathbf{C}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})/2.$$

Proof: Independence is established by showing that $\widetilde{\mathbf{B}}_0$ *is independent of* $\mathbf{E}$*. The proof for the distribution of* $\mathbf{E}$ *is straightforward. To obtain the distribution of* $\mathbf{H}$*, consider the distribution of* $\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-\frac{1}{2}}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M} - \mathbf{\Delta})$*. It is straightforward to show that*

$$\mathrm{vec}\left(\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-\frac{1}{2}}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M} - \mathbf{\Delta})\right) \sim$$

$$\mathrm{N}\left[\mathrm{vec}\left(\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-\frac{1}{2}}(\mathbf{C}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})\right), (\mathbf{M}'\mathbf{\Sigma}\mathbf{M} \otimes \mathbf{I}_q)\right].$$

Thus, the rows of $\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-\frac{1}{2}}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M} - \mathbf{\Delta})$ *are independently and normally distributed with common covariance. The distribution of* $\mathbf{H}$ *follows from a prior theorem.*

□

Corollary 8.8.1: $U \sim U(d, q, n-r, \mathbf{\Lambda})$.

8.4.2 Case II: M is $d \times k$ with rank k

There are several ways to obtain the the LR test of $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ for the case of non-square full column rank $\mathbf{M}$. The easiest way to think about this problem is to transform it into a simpler problem. First, post-multiply the model $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}$ by $\mathbf{M}$ to obtain the reduced model:

$$\mathbf{Y}\mathbf{M} = \mathbf{X}\mathbf{B}\mathbf{M} + \mathbf{U}\mathbf{M},$$

or, equivalently,

$$\mathbf{Z}_1 = \mathbf{X}\mathbf{\Theta}_1 + \mathbf{V}_1,$$

where $\mathbf{Z}_1 = \mathbf{Y}\mathbf{M}$, $\mathbf{\Theta}_1 = \mathbf{B}\mathbf{M}$, and $\mathbf{V}_1 = \mathbf{U}\mathbf{M}$. The distribution of $\mathrm{vec}(\mathbf{Z}_1)$ is

$$\mathrm{vec}(\mathbf{Z}_1) \sim [\mathrm{vec}(\mathbf{X}\mathbf{\Theta}_1), (\mathbf{\Omega}_{11} \otimes \mathbf{I}_n)],$$

where $\mathbf{\Omega}_{11} = \mathbf{M}'\mathbf{\Sigma}\mathbf{M}$. Now use the reduced model and derive the LR test of $\mathrm{H}_0\colon \mathbf{C}'\mathbf{\Theta}_1 = \mathbf{\Delta}$.

Theorem 8.9 *The LR test of* $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *versus* $\mathrm{H}_a\colon \mathbf{C}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta}$ *in the reduced model is to reject* H_0 *for small values of*

$$U = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|},$$

where

$$\mathbf{E} = \mathbf{M}'\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}\mathbf{M},$$

and

$$\mathbf{H} = (\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta}).$$

Proof: Use the previous section to obtain the LR test of $\mathrm{H}_0\colon \mathbf{C}'\mathbf{\Theta}_1\mathbf{M}^* = \mathbf{\Delta}$*, where* $\mathbf{M}^* = \mathbf{I}_k$*. Then simplify.*

□

There is a question that one must ask before using Theorem 8.9. Information is being discarded when one reduced the dimension of the model from d to k. The question is whether or not the discarded information is relevant to the problem. To be specific, let $\mathbf{P}_M$ be a projection operator that projects onto $\mathcal{R}(\mathbf{M})$. The projection operator can be orthogonal or non-orthogonal. The linear function $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_M)$ is ignored in Theorem 8.9. Is there information in $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_M)$ that could be used to improve the test of H_0? The next two theorems show that it is OK to ignore this information. The linear function $\mathbf{Y}(\mathbf{I}_d - \mathbf{P}_M)$ gives no information about whether H_0 is true or false.

Theorem 8.10 derives the LRT of H_0 by using the MLEs of $\mathbf{B}$ and $\mathbf{\Sigma}$ that were derived in Theorem 8.6. Theorem 8.11 starts from scratch and maximizes the likelihood function under H_0 and H_a. They give the same result, so you can pick the one that you like best.

Theorem 8.10 *The LR test of* $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *versus* $\mathrm{H}_a\colon \mathbf{C}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta}$ *in the full model is identical to the LR test of* $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *versus* $\mathrm{H}_a\colon \mathbf{C}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta}$ *in the reduced model. Thus, using the test given in Theorem 8.9 does not entail a loss of efficiency.*

Outline of Proof: The matrix of error sum of squares under H_0 *is*

$$\begin{aligned}(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_0)'(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_0) &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y} \\ &\quad + n^{-1}\widetilde{\mathbf{\Sigma}}_a\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a \\ &= n\widetilde{\mathbf{\Sigma}}_0,\end{aligned}$$

where $\widetilde{\mathbf{B}}_0$ *and* $\widetilde{\mathbf{\Sigma}}_0$ *are given in Theorem 8.6. Accordingly, the LR test statistic is*

$$\begin{aligned}U &= \left[\frac{\max\limits_{\mathrm{H}_0}\exp\{\ell(\mathbf{B},\mathbf{\Sigma}|\mathbf{Y})\}}{\max\limits_{\mathrm{H}_a}\exp\{\ell(\mathbf{B},\mathbf{\Sigma}|\mathbf{Y})\}}\right]^{\frac{2}{n}} = \frac{|\widetilde{\mathbf{\Sigma}}_a|}{|\widetilde{\mathbf{\Sigma}}_0|} \\ &= \frac{|\widetilde{\mathbf{\Sigma}}_a|}{|\widetilde{\mathbf{\Sigma}}_a + n^{-1}\widetilde{\mathbf{\Sigma}}_a\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a|} \\ &= \frac{1}{|\mathbf{I}_d + n^{-1}\mathbf{M}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a|} \\ &= \frac{1}{|\mathbf{I}_s + n^{-1}(\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M})^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})|} \\ &= \frac{|\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M}|}{|\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M} + n^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})|} \\ &= \frac{|n\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M}|}{|n\mathbf{M}'\widetilde{\mathbf{\Sigma}}_a\mathbf{M} + (\widehat{\mathbf{\Delta}} - \mathbf{\Delta})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\widehat{\mathbf{\Delta}} - \mathbf{\Delta})|} \\ &= \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|},\end{aligned}$$

where $\mathbf{E}$ *and* $\mathbf{H}$ *are given in Theorem 8.9.*

□

Theorem 8.11 *The LR test of* $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *versus* $\mathrm{H}_a\colon \mathbf{C}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta}$ *in the full model is identical to the LR test of* $\mathrm{H}_0\colon \mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *versus* $\mathrm{H}_a\colon \mathbf{C}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta}$ *in the reduced model. Thus, using the test given in Theorem 8.9 does not entail a loss of efficiency.*

Outline of Proof: Transform from $\mathbf{Y}$ *to* $\mathbf{Z} = \mathbf{Y}\begin{pmatrix}\mathbf{M} & \mathbf{R}\end{pmatrix}$, *where* $\mathbf{R}$ *is* $d \times (d-k)$ *with rank* $d-k$ *and satisfies* $\mathbf{M}'\mathbf{R} = \mathbf{0}$. *That is, the columns of* $\mathbf{R}$ *form a basis set for* $\mathcal{N}(\mathbf{M}')$. *Let*

$$\mathbf{Z} = \begin{pmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{pmatrix}, \text{ where } \mathbf{Z}_1 = \mathbf{YM} \text{ and } \mathbf{Z}_2 = \mathbf{YR}.$$

Factor the joint pdf of $\mathbf{Z}$ *as*

$$f_Z(\mathbf{Z}_1, \mathbf{Z}_2) = f_{Z_1}(\mathbf{Z}_1) \times f_{Z_2|Z_1}(\mathbf{Z}_2|\mathbf{Z}_1).$$

Let

$$\mathbf{B}\begin{pmatrix}\mathbf{M} & \mathbf{R}\end{pmatrix} = \begin{pmatrix}\mathbf{\Theta}_1 & \mathbf{\Theta}_2\end{pmatrix}, \text{ where } \mathbf{\Theta}_1 = \mathbf{BM}, \text{ and } \mathbf{\Theta}_2 = \mathbf{BR}.$$

The marginal model for $\mathbf{Z}_1$ *is*

$$\text{vec}(\mathbf{Z}_1) \sim \text{N}\left[\text{vec}(\mathbf{X\Theta}_1), \mathbf{\Omega}_{11} \otimes \mathbf{I}_n\right],$$

where $\mathbf{\Omega}_{11} = \mathbf{M}'\mathbf{\Sigma}\mathbf{M}$. *The conditional model for* $\mathbf{Z}_2$ *given* $\mathbf{Z}_1$ *is*

$$\text{vec}(\mathbf{Z}_2)|\mathbf{Z}_1 \sim \text{N}\left[\text{vec}(\mathbf{X\Theta}_{2\cdot1} + \mathbf{Z}_1\mathbf{\Gamma}), \mathbf{\Omega}_{22\cdot1} \otimes \mathbf{I}_n\right],$$

where

$$\mathbf{\Omega} = \begin{pmatrix}\mathbf{M}'\mathbf{\Sigma}\mathbf{M} & \mathbf{M}'\mathbf{\Sigma}\mathbf{R} \\ \mathbf{R}'\mathbf{\Sigma}\mathbf{M} & \mathbf{R}'\mathbf{\Sigma}\mathbf{R}\end{pmatrix}, \quad \mathbf{\Theta}_{2\cdot1} = \mathbf{\Theta}_2 - \mathbf{\Theta}_1\mathbf{\Gamma}, \text{ and } \mathbf{\Gamma} = \mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}.$$

The key to obtaining the LR test is to notice that under $\text{H}_0\colon \mathbf{C}'\mathbf{BM} = \mathbf{\Delta}$, *the MLE's of* $\mathbf{\Theta}_1$ *and* $\mathbf{\Omega}_{11}$ *are restricted, but the MLE's of* $\mathbf{\Theta}_{2\cdot1}$, $\mathbf{\Gamma}$, *and* $\mathbf{\Omega}_{22\cdot1}$ *are not restricted. Thus, the latter parameters have the same MLE's under* H_0 *and* H_a. *Accordingly, the maximized conditional pdf* $[F_{Z_2|Z_1}]$ *is identical in the numerator and denominator of the LR test statistic. After cancellation, the LR test statistic simplifies to the ratio of the maximized marginal pdfs. This is the test described in Theorem 8.9.*

□

Note that the form of the test statistic is identical regardless of whether $\mathbf{M}$ is square and non-singular or non-square. To obtain critical values of the above LR statistic, the following result can be used.

Theorem 8.12 *The matrices* $\mathbf{E}$ *and* $\mathbf{H}$ *are independently distributed as*

$$\mathbf{E} \sim \text{W}_k(n-r, \mathbf{M}'\mathbf{\Sigma}\mathbf{M}, \mathbf{0})$$

and

$$\mathbf{H} \sim \text{W}_k(q, \mathbf{M}'\mathbf{\Sigma}\mathbf{M}, \mathbf{\Lambda}),$$

where

$$\mathbf{\Lambda} = (\mathbf{M}'\mathbf{\Sigma}\mathbf{M})^{-1}(\mathbf{C}'\mathbf{BM} - \mathbf{\Delta})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}]^{-1}(\mathbf{C}'\mathbf{BM} - \mathbf{\Delta})/2.$$

Proof: Use prior results.

□

Corollary 8.12.1: $U \sim U(k, q, n-r, \mathbf{\Lambda})$.

8.5 ALTERNATIVE TEST CRITERIA

The other common multivariate criteria are obtained by considering alternative functions of the characteristic roots of $\mathbf{E}^{-1}\mathbf{H}$:

1. Roy's Max root: $\varphi_{\max} = r_1(\mathbf{E}^{-1}\mathbf{H})$ or $\theta_{\max} = r_1[\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}]$.
2. Lawley-Hotelling trace: $T_g^2 = (n-r)\,\text{tr}(\mathbf{E}^{-1}\mathbf{H})$.
3. Pillai trace: $V = \text{tr}[\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}]$.

8.6 SIMULTANEOUS INFERENCE BASED ON THE UI PRINCIPLE

The goal in this Section is to construct simultaneous confidence intervals on linear functions of $\mathbf{L}'\mathbf{B}\mathbf{M}$.

8.6.1 The Union Intersection Test

Consider testing H: $\mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t} = \mathbf{f}'\mathbf{\Delta}\mathbf{t}$ for some specified $\mathbf{f}: q \times 1$ and $\mathbf{t}: k \times 1$. Note that

$$\mathrm{H}_0\colon \mathbf{L}'\mathbf{B}\mathbf{M} = \mathbf{\Delta} \Longleftrightarrow \bigcap_{\mathbf{f},\mathbf{t}} \mathrm{H}\colon \mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t} = \mathbf{f}'\mathbf{\Delta}\mathbf{t}$$

and

$$\mathrm{H}_a\colon \mathbf{L}'\mathbf{B}\mathbf{M} \neq \mathbf{\Delta} \Longleftrightarrow \bigcup_{\mathbf{f},\mathbf{g}} \mathrm{H}\colon \mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t} \neq \mathbf{f}'\mathbf{\Delta}\mathbf{t}.$$

To obtain a UI test of H_0, consider the univariate test of H for fixed $\mathbf{f}$ and $\mathbf{t}$.

To construct a test of H: $\mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t} = \mathbf{f}'\mathbf{\Delta}\mathbf{t}$, note that

$$\mathbf{f}'(\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})\mathbf{t} \sim \mathrm{N}\left[\mathbf{f}'(\mathbf{L}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})\mathbf{t},\ \mathbf{t}'\mathbf{M}'\mathbf{\Sigma}\mathbf{M}\mathbf{t} \times \mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}\mathbf{f}\right].$$

Thus, for fixed $\mathbf{f}$ and $\mathbf{t}$,

$$Q(\mathbf{f},\mathbf{t}) = \frac{\left[\mathbf{f}'(\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})\mathbf{t}\right]^2}{\mathbf{t}'\mathbf{M}'\mathbf{S}\mathbf{M}\mathbf{t} \times \mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}\mathbf{f}} \sim F_{1,n-r,\lambda},$$

where

$$\mathbf{S} = (n-r)^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y} \text{ and } \lambda = \frac{\left[\mathbf{f}'(\mathbf{L}'\mathbf{B}\mathbf{M} - \mathbf{\Delta})\mathbf{t}\right]^2}{2\mathbf{t}'\mathbf{M}'\mathbf{\Sigma}\mathbf{M}\mathbf{t} \times \mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}\mathbf{f}}.$$

If the coefficient vectors $\mathbf{M}\mathbf{t}$ and $\mathbf{L}\mathbf{f}$ are chosen a a priori, then the LR test of $\mathrm{H}_0\colon \mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t} = 0$ against $\mathrm{H}_a\colon \mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t} \neq 0$ is to reject H_0 if $Q(\mathbf{f},\mathbf{t}) > F^{1-\alpha}_{1,n-r}$.

Theorem 8.13 *Roy's maximum root criterion is a union intersection test statistic for* H_0. *In particular, the UI statistic for testing* $\mathrm{H}_0\colon \mathbf{L}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *is the maximum of* $Q(\mathbf{f},\mathbf{t})$ *over all* $\mathbf{f}$ *and* $\mathbf{t}$. *That is,*

$$Q = \sup_{\mathbf{f},\mathbf{t}} Q(\mathbf{f},\mathbf{t}) = (n-r)\,\mathrm{eval}_1\left(\mathbf{E}^{-1}\mathbf{H}\right),$$

where

$$\mathbf{E} = \mathbf{M}'\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}\mathbf{M}; \quad \mathbf{H} = (\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta})'\left[\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}\right]^{-1}(\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{\Delta}),$$

and $\mathrm{eval}_1(\mathbf{W})$ *is the maximum eigenvalue of* $\mathbf{W}$.

Proof HW or in class.

□

8.6.2 Simultaneous Confidence Intervals

Denote the linear function $\mathbf{f}'\mathbf{L}'\mathbf{B}\mathbf{M}\mathbf{t}$ by ψ and its estimator $\mathbf{f}'\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M}\mathbf{t}$ by $\hat{\psi}$. We wish to construct simultaneous $(1-\alpha)100\%$ confidence intervals for the set of all ψ. That is, we want to construct a set of confidence intervals for all possible ψ such that with probability $1-\alpha$, all of the intervals capture the unknown function. We will use the pivotal quantity method. Let

$$Q^*(\mathbf{f},\mathbf{t}) = \frac{\left[\mathbf{f}'(\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M} - \mathbf{L}'\mathbf{B}\mathbf{M})\mathbf{t}\right]^2}{\mathbf{t}'\mathbf{M}'\mathbf{S}\mathbf{M}\mathbf{t} \times \mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}\mathbf{f}} = \frac{(\hat{\psi} - \psi)^2}{\mathbf{t}'\mathbf{M}'\mathbf{S}\mathbf{M}\mathbf{t} \times \mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}\mathbf{f}}.$$

The quantity Q^* is a pivotal quantity because it depends on the unknown parameter ψ and its distribution is free of ψ:

$$Q^*(\mathbf{f},\mathbf{t}) \sim F_{1,n-r,0}.$$

From the foregoing results, it can be shown that

$$\sup_{\mathbf{f},\mathbf{t}} Q^*(\mathbf{f},\mathbf{t}) = (n-r)\,\mathrm{eval}_1\left(\mathbf{E}^{-1}\mathbf{H}^*\right),$$

where

$$\begin{aligned}\mathbf{E} &= \mathbf{M}'\mathbf{Y}'(\mathbf{I}-\mathbf{H}_x)\mathbf{Y}\mathbf{M};\ \text{and}\\ \mathbf{H}^* &= (\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M}-\mathbf{L}'\mathbf{B}\mathbf{M})'[\mathbf{L}'(\mathbf{X}'\mathbf{X})^-\mathbf{L}]^{-1}(\mathbf{L}'\widetilde{\mathbf{B}}_a\mathbf{M}-\mathbf{L}'\widetilde{\mathbf{B}}\mathbf{M}).\end{aligned}$$

Note that $\mathbf{H}^*$ is independent of $\mathbf{E}$ and has distribution

$$\mathbf{H}^* \sim \mathrm{W}_k(q, \mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}, \mathbf{0}).$$

Let $r^{1-\alpha}_{k,m_H,m_E}$ denote the $(1-\alpha)100$ percentile of the null distribution of

$$\mathrm{eval}_1\left[\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}\right],$$

where $\mathbf{E}$ and $\mathbf{H}$ are independently distributed as $\mathbf{H} \sim \mathrm{W}_k(m_H, \mathbf{I}_s)$ and $\mathbf{E} \sim \mathrm{W}_k(m_E, \mathbf{I}_s)$. Percentiles, $r^{1-\alpha}_{k,m_H,m_E}$ are tabled in Rencher (2002, Table A.10). Percentiles also can be obtained from the charts of Heck (*Annals of Mathematical Statistics*, 1960, **31**, 625–642). For the special case of $m_H = 1$, the percentile simplifies to

$$r^{1-\alpha}_{k,1,m_E} = \frac{kF^{1-\alpha}_{k,m_E-k+1}}{m_E-k+1+kF^{1-\alpha}_{k,m_E-k+1}}.$$

The $(1-\alpha)100$ percentile of $m_E\,\mathrm{eval}_1\left(\mathbf{E}^{-1}\mathbf{H}\right)$ is $\ell^{1-\alpha}_{k,m_H,m_E}$ where

$$\ell^{1-\alpha}_{k,m_H,m_E} = \frac{m_E r^{1-\alpha}_{k,m_H,m_E}}{1-r^{1-\alpha}_{k,m_H,m_E}}.$$

If $m_h = 1$, then,

$$\ell^{1-\alpha}_{k,1,m_E} = \frac{m_E kF^{1-\alpha}_{k,m_E-k+1}}{m_E-k+1}.$$

Accordingly, simultaneous confidence intervals are obtained from

$$\begin{aligned}&\Pr\left\{(n-r)\,\mathrm{eval}_1\left(\mathbf{E}^{-1}\mathbf{H}^*\right) \le \ell^{1-\alpha}_{k,q,n-r}\right\} = 1-\alpha\\ &\Longrightarrow \Pr\left[\sup_{\mathbf{f},\mathbf{t}} Q^*(\mathbf{f},\mathbf{t}) \le \ell^{1-\alpha}_{k,q,n-r}\right] = 1-\alpha\\ &\Longrightarrow \Pr\left[\frac{(\hat{\psi}-\psi)^2}{\mathbf{t}'\mathbf{M}'\mathbf{S}\mathbf{M}\mathbf{t}\times\mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^-\mathbf{L}\mathbf{f}} \le \ell^{1-\alpha}_{k,q,n-r}\ \forall\ \mathbf{f},\mathbf{t}\right] = 1-\alpha\\ &\Longrightarrow \mathrm{pr}\left[\hat{\psi}-g \le \psi \le \hat{\psi}+g\ \forall\ \mathbf{f},\mathbf{t}\right] = 1-\alpha,\end{aligned}$$

where

$$g = \sqrt{\mathbf{t}'\mathbf{M}'\mathbf{S}\mathbf{M}\mathbf{t}\times\mathbf{f}'\mathbf{L}'(\mathbf{X}'\mathbf{X})^-\mathbf{L}\mathbf{f}\times\ell^{1-\alpha}_{k,q,n-r}}$$

and $\mathbf{S} = \mathbf{Y}'(\mathbf{I}-\mathbf{H}_x\mathbf{Y}/(n-r)$.

A simultaneous critical value for the "F" statistic corresponding to $\mathrm{H}_0\colon \mathbf{l}'\mathbf{B}\mathbf{m} = \delta$ is $\ell^{1-\alpha}_{k,q,n-r}$, where $\mathbf{l} = \mathbf{L}\mathbf{f}$ for some $\mathbf{f}$ and $\mathbf{m} = \mathbf{M}\mathbf{t}$ for some $\mathbf{t}$.

8.7 ANALYSIS OF REPEATED MEASURES

8.7.1 Univariate Versus Multivariate Analyses

8.7.2 k Group Profile Analysis

8.8 ANALYSIS OF GROWTH CURVES

8.8.1 Introduction

Growth curve models constitute an important general class of repeated measures models. The usual repeated measures model is a special case of the growth curve model. The model can be written as follows:

$$\mathbf{Y} = \mathbf{XBG} + \mathbf{W}, \tag{8.1}$$

where $\mathbf{Y}$ is a $n \times d$ random matrix whose entries are d repeated measures on each of n cases; $\mathbf{X}$ is an $n \times p$ known design matrix with rank r; $\mathbf{B}$ is a $p \times b$ matrix of location parameter; $\mathbf{G}$ is a $b \times d$ known design matrix with rank $g < d$; and $\text{vec}(\mathbf{W}) \sim \text{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}_n)$. The matrix $\mathbf{X}$ is called a between-subjects design matrix whereas the matrix $\mathbf{G}$ is called a within-subjects design matrix.

The model in (8.1) is called a generalized MANOVA or GMANOVA model. It was proposed by Potthoff and Roy (1964). Comprehensive reviews of the GMANOVA model were given by D. von Rosen (1991, The Growth Curve Model: A Review, *Communications in Statistics — Theory and Methods*, **20**, 2791–2822) and A. M. Kshirsagar and W. B. Smith (1995, *Growth Curves*, New York: Marcel Dekker). This section describes the conditional approach to analyzing growth curves which was pioneered by C. R. Rao (1966, Covariance Adjustment and Related Problems in Multivariate Analysis, in P. R. Krishnaiah, ed., *Multivariate Analysis*, pp. 87–103, New York: Academic Press) and C. G. Khatri (1966, A Note on a MANOVA Model Applied to Problems in Growth Curve, *Annals of the Institute of Statistical Mathematics*, **18**, 75–86).

Typically, the rows of $\mathbf{G}$ consist of polynomial functions of time. For example, suppose that observations are taken at six occasions on two groups of subjects. The observations are taken at $1, 2, 5, 7, 9$, and 12 time units after the start of the study. Further, suppose that it is believed that all changes over time can be summarized in term of linear, quadratic, and cubic functions. There are several possible parameterizations for this growth model. One possibility is to write the $d \times 1$ vector of expected responses in the j^{th} group as

$$\text{E}(\overline{\mathbf{y}}_j) = \alpha_{j,0}\mathbf{1}_d + \alpha_{j,1}(\mathbf{t} - \mathbf{1}_d\overline{t}) + \alpha_{j,2}(\mathbf{t} - \mathbf{1}_d\overline{t})^{(2)} + \alpha_{j,3}(\mathbf{t} - \mathbf{1}_d\overline{t})^{(3)},$$

where $\mathbf{t} = \begin{pmatrix} 1 & 2 & 5 & 7 & 9 & 12 \end{pmatrix}'$; $\overline{t} = d^{-1}\mathbf{1}_d'\mathbf{t} = 6$; and $(\mathbf{t} - \mathbf{1}_d\overline{t})^{(h)}$ is the $d \times 1$ vector obtained by raising each entry of $\mathbf{t} - \mathbf{1}_d\overline{t}$ to the h^{th} power. For this example

$$\mathbf{G} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -5 & -4 & -1 & 1 & 3 & 6 \\ 25 & 16 & 1 & 1 & 9 & 36 \\ -125 & -64 & -1 & 1 & 27 & 216 \end{pmatrix}.$$

Continuing with the example, suppose that the coefficients are

$$\mathbf{A} = \begin{pmatrix} \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \end{pmatrix} = \begin{pmatrix} 10.0 & 0.9 & 0.6 & 0.3 \\ 20.0 & 1.2 & 0.8 & 0.0 \end{pmatrix}$$

and that the between subjects design matrix is

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \bigoplus_{j=1}^{2} \mathbf{1}_{n_j} \end{pmatrix}.$$

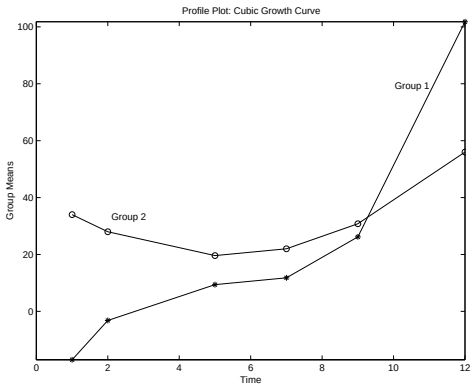
One of the infinite possible values for $\mathbf{B}$ that will generate the matrix $\mathbf{A} = \{\alpha_{j,h}\}$ is

$$\mathbf{B} = \begin{pmatrix} 4.0 & 4.0 & 0.0 & 4.0 \\ 6.0 & -3.1 & 0.6 & -3.7 \\ 16.0 & -2.8 & 0.8 & -4.0 \end{pmatrix}.$$

The first column of $\mathbf{B}$ contains intercept parameters, the second column contains linear slope parameters, etc. The mean response profiles for this example are

$$\mathrm{E}(\overline{\mathbf{Y}}) = \begin{pmatrix} -17.0 & -3.2 & 9.4 & 11.8 & 26.2 & 101.8 \\ 34.0 & 28.0 & 19.6 & 22.0 & 30.8 & 56.0 \end{pmatrix}.$$

A profile plot of the expected means appears below.



A second possible parameterization is to use orthogonal polynomials. In this example, the matrix, $\mathbf{G}$, containing the coefficients of orthogonal polynomials is

$$\mathbf{G} = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ -5.0000 & -4.0000 & -1.0000 & 1.0000 & 3.0000 & 6.0000 \\ 13.4015 & 3.7879 & -13.0530 & -14.2803 & -7.5076 & 17.6515 \\ -29.7999 & 22.8847 & 39.2613 & -7.8853 & -45.4835 & 21.0227 \end{pmatrix}.$$

The first row of $\mathbf{G}$ represents the intercept, the second row consists of linear coefficients, etc. The matrix $\mathbf{G}$ can be computed by using the following MATLAB program:

```
b=4;
t=[1 2 5 7 9 12]';
d=length(t);
G=ones(1,d);
for i=1:b-1
ti=(eye(d)-ppo(G'))*t.^i;
G=[G; ti'];
end;
```

For this example, one of the infinite possible values for $\mathbf{B}$ that will generate the matrix $\mathbf{A} = \{\alpha_{j,h}\}$ is

$$\mathbf{B} = \begin{pmatrix} 17.7444 & 3.5545 & 0.6473 & 0.1000 \\ 3.7556 & 5.4182 & 0.4946 & 0.2000 \\ 13.9889 & -1.8636 & 0.1527 & -0.1000 \end{pmatrix}.$$

The expected responses are the same as those given earlier.

8.8.2 Parameter Estimation

8.8.2.1 *Estimability*

Consider the growth curve model in (8.1). Suppose that an unbiased estimator of $\mathbf{C}'\mathbf{B}\mathbf{M}$ is desired, where $\mathbf{C}$ is a known $p \times q$ matrix and $\mathbf{M}$ is a known $b \times k$ matrix. The first question that must be answered is — under what conditions is $\mathbf{C}'\mathbf{B}\mathbf{M}$ estimable?

Lemma 8.1 *Let* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, *and* $\mathbf{D}$ *be any matrices such that* $\mathbf{A}$ *and* $\mathbf{C}$ *have the same dimensions;* $\mathbf{B}$ *and* $\mathbf{D}$ *have the same dimensions; and each matrix contains at least one non-zero entry. Then*

$$\mathbf{A} \otimes \mathbf{B} = \mathbf{C} \otimes \mathbf{D} \Longleftrightarrow \mathbf{A} = \alpha\mathbf{C} \text{ and } \mathbf{B} = \alpha^{-1}\mathbf{D},$$

where α *is a non-zero scalar.*

Outline of proof: use the definition of Kronecker multiplication.

□

Theorem 8.14 (Estimability in Growth Curve Models) *If each of* $\mathbf{C}$ *and* $\mathbf{M}$ *have at least one non-zero entry, then* $\mathbf{C}'\mathbf{B}\mathbf{M}$ *is estimable if and only if* $\mathbf{C} \in \mathcal{R}(\mathbf{X}')$ *and* $\mathbf{M} \in \mathcal{R}(\mathbf{G})$.

Outline of proof: Let $\psi = \text{vec}(\mathbf{C}'\mathbf{B}\mathbf{M})$ *and let* $\widehat{\psi} = \mathbf{L}'\mathbf{y} + \mathbf{h}$, *where* $\mathbf{y} = \text{vec}(\mathbf{Y})$ *and* $\mathbf{h}$ *is a vector of constants. First, verify that estimability requires that* $\mathbf{h} = \mathbf{0}$ *and that* $(\mathbf{M} \otimes \mathbf{C}) \in \mathcal{R}(\mathbf{G} \otimes \mathbf{X}')$. *Second, use PPOs and the above Lemma to verify that* $\mathbf{C} \in \mathcal{R}(\mathbf{X}')$ *and* $\mathbf{M} \in \mathcal{R}(\mathbf{G})$ *are necessary conditions for estimability. To verify that they are sufficient conditions, assume that* $\mathbf{C} \in \mathcal{R}(\mathbf{X}')$ *and* $\mathbf{M} \in \mathcal{R}(\mathbf{G})$ *are both satisfied and examine the expectation of* $\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\mathbf{G}'(\mathbf{G}\mathbf{G}')^{-}\mathbf{M}$.

□

Theorem 8.15 (BLUE When $\boldsymbol{\Sigma}$ is Known) *If* $\boldsymbol{\Sigma}$ *is known and* $\boldsymbol{\Psi} = \mathbf{C}'\mathbf{B}\mathbf{M}$ *is estimable, then the BLUE of* $\boldsymbol{\Psi}$ *is*

$$\widehat{\boldsymbol{\Psi}} = \mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M}, \text{ where } \widetilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}\boldsymbol{\Sigma}^{-1}\mathbf{G}'(\mathbf{G}\boldsymbol{\Sigma}^{-1}\mathbf{G}')^{-}.$$

Proof: HW.

□

Theorem 8.16 (MLE When $\boldsymbol{\Sigma}$ is Not Known) *If* $\boldsymbol{\Sigma}$ *is unknown then the MLEs of* $\mathbf{B}$ *and* $\boldsymbol{\Sigma}$ *are*

$$\widetilde{\mathbf{B}}_G = \widetilde{\mathbf{B}}\widetilde{\boldsymbol{\Sigma}}^{-1}\mathbf{G}'(\mathbf{G}\widetilde{\boldsymbol{\Sigma}}^{-1}\mathbf{G}')^{-} \text{ and } \widetilde{\boldsymbol{\Sigma}}_G = \widetilde{\boldsymbol{\Sigma}} + n^{-1}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})'\mathbf{X}'\mathbf{X}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})$$

where

$$\widetilde{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}; \quad \widetilde{\boldsymbol{\Sigma}} = n^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}; \text{ and } \mathbf{H}_x = \text{ppo}(\mathbf{X}).$$

Outline of Proof: Write $\mathbf{G}$ *in terms of its singular values and vectors:* $\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{V}'$. *Let* $\mathbf{R} = \begin{pmatrix}\mathbf{R}_1 & \mathbf{R}_2\end{pmatrix}$, *where*

$$\mathbf{R}_1 = \mathbf{T}^{-1}\mathbf{V}(\mathbf{V}'\mathbf{T}^{-1}\mathbf{V})^{-1}\mathbf{D}^{-1}; \quad \mathbf{A}\mathbf{R}_2' = \mathbf{I}_d - \mathbf{V}(\mathbf{V}'\mathbf{T}^{-1}\mathbf{V})^{-1}\mathbf{V}'\mathbf{T}^{-1};$$

$\mathbf{T}$ *is an arbitrary positive definite matrix of dimension* $d \times d$; *and each of the matrices in the factorization* $\mathbf{A}\mathbf{R}_2'$ *has full column rank. The dimensions of* $\mathbf{R}_1$ *and* $\mathbf{R}_2$ *are* $d \times g$ *and* $d \times (d - g)$ *respectively. Equivalent representations for* $\mathbf{R}_1$ *and* $\mathbf{R}_2$ *are*

$$\mathbf{R}_1 = \mathbf{P}'_{\mathbf{G}',\mathbf{T}}\mathbf{V}\mathbf{D}^{-1} \text{ and } \mathbf{A}\mathbf{R}_2' = \mathbf{I}_d - \mathbf{P}_{\mathbf{G}',\mathbf{T}}$$

where

$$\mathbf{P}_{\mathbf{G}',\mathbf{T}} = \mathbf{G}'(\mathbf{G}\mathbf{T}^{-1}\mathbf{G}')^{-}\mathbf{G}\mathbf{T}^{-1}.$$

Note that $\mathbf{G}\mathbf{R}_1 = \mathbf{U}$ *and* $\mathbf{G}\mathbf{R}_2 = \mathbf{0}$.

Write the log likelihood as

$$\begin{aligned}\ell(\mathbf{B}, \boldsymbol{\Sigma}|\mathbf{Y}) &= -\frac{1}{2}\text{tr}\left[(\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{G})'(\mathbf{Y} - \mathbf{X}\mathbf{B}\mathbf{G})\boldsymbol{\Sigma}^{-1}\right] - \frac{n}{2}\ln|\boldsymbol{\Sigma}| \\ &= -\frac{1}{2}\text{tr}\left[(\mathbf{Z} - \mathbf{X}\mathbf{B}\mathbf{G}\mathbf{R})'(\mathbf{Z} - \mathbf{X}\mathbf{B}\mathbf{G}\mathbf{R})\boldsymbol{\Omega}^{-1}\right] - \frac{n}{2}\ln|\boldsymbol{\Omega}|\end{aligned}$$

plus constants, where

$$\mathbf{Z} = \mathbf{Y}\mathbf{R} = \begin{pmatrix}\mathbf{Y}\mathbf{R}_1 & \mathbf{Y}\mathbf{R}_2\end{pmatrix}$$

and

$$\boldsymbol{\Omega} = \mathbf{R}'\boldsymbol{\Sigma}\mathbf{R} = \begin{pmatrix}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22}\end{pmatrix}.$$

Factor the log likelihood as

$$\ell(\mathbf{B},\mathbf{\Sigma}|\mathbf{Y}) = \ell_2(\mathbf{\Omega}_{22}|\mathbf{Z}_2) + \ell_{1\cdot 2}(\mathbf{B},\mathbf{\Gamma},\mathbf{\Omega}_{11\cdot 2}|\mathbf{Z}_1)$$

where

$$\ell_2(\mathbf{\Omega}_{22}|\mathbf{Z}_2) = -\frac{1}{2}\operatorname{tr}\left[\mathbf{Z}_2'\mathbf{Z}_2\mathbf{\Omega}_{22}^{-1}\right] - \frac{n}{2}\ln|\mathbf{\Omega}_{22}|;$$

$$\ell_{1\cdot 2}(\mathbf{B},\mathbf{\Gamma},\mathbf{\Omega}_{11\cdot 2}|\mathbf{Z}_1) =$$

$$-\frac{1}{2}\operatorname{tr}\left[(\mathbf{Z}_1 - \mathbf{XBU} - \mathbf{Z}_2\mathbf{\Gamma})'(\mathbf{Z}_1 - \mathbf{XBU} - \mathbf{Z}_2\mathbf{\Gamma})\mathbf{\Omega}_{11\cdot 2}^{-1}\right] - \frac{n}{2}\ln|\mathbf{\Omega}_{11\cdot 2}|;$$

and $\mathbf{\Gamma} = \mathbf{\Omega}_{22}^{-1}\mathbf{\Omega}_{21}$. *To complete the proof, maximize the two components of the log likelihood separately, let* $\mathbf{T} = \widetilde{\mathbf{\Sigma}}$ *and compute* $\widetilde{\mathbf{\Sigma}}_G$ *as*

$$\widetilde{\mathbf{\Sigma}}_G = n^{-1}(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_G\mathbf{G})'(\mathbf{Y} - \mathbf{X}\widetilde{\mathbf{B}}_G\mathbf{G}).$$

□

8.8.3 Hypothesis tests and Confidence Intervals

One question of interest is whether the growth curve model is adequate. That is, does $\mathbf{G}$ contain the correct polynomial functions. This question may be answered by testing $\mathrm{H}_0\colon \mathrm{E}(\mathbf{Y}) = \mathbf{XBG}$ against $\mathrm{H}_a\colon \mathbf{E}(\mathbf{Y}) = \mathbf{XB}$. Note that $\mathbf{B}$ is a $p \times b$ matrix under H_0 and a $p \times d$ matrix under H_a. This test is equivalent to testing $\mathrm{H}_0\colon \mathbf{C}'\mathbf{BM} = \mathbf{0}$ in the usual repeated measures model, where $\mathbf{C}$ is a $p \times r$ matrix that satisfies $\mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{X}')$ and $\mathbf{M}$ is a $d \times (d-g)$ matrix with rank $d-g$ that satisfies $\mathbf{GM} = \mathbf{0}$.

Theorem 8.17 (Test of Lack of Fit of GMANOVA Model) *Let* $\mathbf{C}$ *be a* $p \times r$ *matrix whose columns span* $\mathcal{R}(\mathbf{X}')$ *and let* $\mathbf{M}$ *be a* $d \times (d-g)$ *matrix whose columns span* $\mathcal{N}(\mathbf{G})$. *Then the LRT test of* $\mathrm{H}_0\colon \mathrm{E}(\mathbf{Y}) = \mathbf{XBG}$ *against* $\mathrm{H}_a\colon \mathbf{E}(\mathbf{Y}) = \mathbf{XB}$ *is to reject* H_0 *for small values of* $U = |\mathbf{E}|/|\mathbf{E}+\mathbf{H}|$, *where*

$$\mathbf{H} = (\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}\mathbf{M}) \text{ and } \mathbf{E} = \mathbf{M}'\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{YM}.$$

These matrices are distributed independently as

$$\mathbf{H} \sim \mathrm{W}_{d-g}(r, \mathbf{M}'\mathbf{\Sigma}\mathbf{M}, \mathbf{\Lambda}) \text{ and } \mathbf{E} \sim \mathrm{W}_{d-g}(n-r, \mathbf{M}'\mathbf{\Sigma}\mathbf{M}),$$

where

$$\mathbf{\Lambda} = (\mathbf{M}'\mathbf{\Sigma}\mathbf{M})^{-1}(\mathbf{C}'\mathbf{BM})'\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\right]^{-1}(\mathbf{C}'\mathbf{BM})/2.$$

Proof: Use Theorems 8.9 and 8.10.

□

Theorem 8.18 *The non-zero eigenvalues of* $\mathbf{E}^{-1}\mathbf{H}$ *are identical to the non-zero eigenvalues of* $\mathbf{E}_*^{-1}\mathbf{H}_*$, *where* $\mathbf{E}$ *and* $\mathbf{H}$ *are defined in Theorem (8.17);*

$$\mathbf{E}_* = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}; \text{ and } \mathbf{H}_* = (\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})'\mathbf{X}'\mathbf{X}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G}).$$

Outline of Proof: Let $\mathbf{M} = \mathbf{R}_2$, *where* $\mathbf{AR}_2' = \mathbf{I}_d - \mathbf{P}_{\mathbf{G}',\widetilde{\mathbf{\Sigma}}}$ *and use properties of projection matrices to verify that* $\mathbf{R}_2'\widetilde{\mathbf{\Sigma}}\mathbf{R}_2 = (\mathbf{A}'\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{A})^{-1}$.

□

The following procedures can be used to construct confidence intervals and hypothesis tests about $\mathbf{C}'\mathbf{BM}$ in the GMANOVA model.

Theorem 8.19 *Suppose that* $\mathbf{\Psi} = \mathbf{C}'\mathbf{BM}$ *is estimable in the GMANOVA model. Then the MLE of* $\mathbf{\Psi}$ *is*

$$\widehat{\mathbf{\Psi}} = \mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M}$$

and the distribution of $\operatorname{vec}(\widehat{\mathbf{\Psi}})$, *conditional on* $\mathbf{Z}_2$ *is*

$$\operatorname{vec}(\widehat{\mathbf{\Psi}})|\mathbf{Z}_2 \sim \mathrm{N}\left[\operatorname{vec}(\mathbf{\Psi}), \mathbf{\Theta}\right],$$

where

$$\mathbf{\Theta} = \mathbf{M}'\mathbf{U}\mathbf{\Omega}_{11\cdot 2}\mathbf{U}'\mathbf{M} \otimes \left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C} + n^{-1}\mathbf{C}'(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})\widetilde{\mathbf{\Sigma}}^{-1}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})'\mathbf{C}\right],$$

where $\widetilde{\mathbf{\Sigma}} = n^{-1}\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}$ *and* $\mathbf{H}_x = \text{ppo}(\mathbf{X})$. *Furthermore, the MLE of* $\mathbf{\Theta}$ *is*

$$\widetilde{\text{Disp}}(\widehat{\mathbf{\Psi}}) = \widetilde{\mathbf{\Theta}} =$$

$$\mathbf{M}'(\mathbf{G}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{G}')^{-}\mathbf{M} \otimes \left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C} + n^{-1}\mathbf{C}'(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})\widetilde{\mathbf{\Sigma}}^{-1}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})'\mathbf{C}\right],$$

and an unbiased estimator of $\mathbf{\Theta}$ *is*

$$\widehat{\mathbf{\Theta}} = \frac{n}{n - r - d + g}\widetilde{\mathbf{\Theta}}.$$

Outline of Proof: First, verify that the maximizer of $\ell_{1\cdot 2}(\mathbf{B}, \mathbf{\Omega}_{11\cdot 2}, \mathbf{\Gamma}|\mathbf{Z}_1)$ *with respect to* $\mathbf{B}$ *is*

$$\widetilde{\mathbf{B}}_G = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}_1\mathbf{U}' - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Z}_2[\mathbf{Z}_2'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Z}_2]^{-1}\mathbf{Z}_2'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Z}_1\mathbf{U}'.$$

Second, compute the dispersion of $\mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M}$ *conditional on* $\mathbf{Z}_2$. *Note,* $\mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M}$ *is a linear function of* $\mathbf{Z}_1$ *when conditioning on* $\mathbf{Z}_2$. *Finally, let* $\mathbf{T} = \widetilde{\mathbf{\Sigma}}$ *and simplify.*

□

Theorem 8.20 (LRT in GMANOVA Model) *Suppose that* $\mathbf{\Psi} = \mathbf{C}'\mathbf{B}\mathbf{M}$ *is estimable in the GMANOVA model, where* $\mathbf{C}$ *is* $q \times p$ *with rank* q *and* $\mathbf{M}$ *is* $db \times k$ *with rank* k. *Then, the likelihood ratio test of* $\mathbf{C}'\mathbf{B}\mathbf{M} = \mathbf{\Delta}$ *in the GMANOVA model is to reject* H_0 *for small values of* $U = |\mathbf{E}|/|\mathbf{E} + \mathbf{H}|$, *where*

$$\mathbf{E} = n\mathbf{M}'(\mathbf{G}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{G}')^{-}\mathbf{M} \text{ and}$$

$$\mathbf{H} = (\mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M} - \mathbf{\Delta})'\times$$

$$\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C} + n^{-1}\mathbf{C}'(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})\widetilde{\mathbf{\Sigma}}^{-1}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})'\mathbf{C}\right]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M} - \mathbf{\Delta}).$$

Furthermore, conditional on $\mathbf{Z}_2$, $\mathbf{E}$ *and* $\mathbf{H}$ *are independently distributed as*

$$\mathbf{E} \sim \text{W}_k(n - r - d + g, \mathbf{M}'\mathbf{U}\mathbf{\Omega}_{11\cdot 2}\mathbf{U}'\mathbf{M}) \text{ and}$$

$$\mathbf{H} \sim \text{W}_k(q, \mathbf{M}'\mathbf{U}\mathbf{\Omega}_{11\cdot 2}\mathbf{U}'\mathbf{M}, \mathbf{\Lambda}),$$

where

$$\mathbf{\Lambda} = (\mathbf{M}'\mathbf{U}\mathbf{\Omega}_{11\cdot 2}\mathbf{U}'\mathbf{M})^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M} - \mathbf{\Delta})'\times$$

$$\left[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C} + n^{-1}\mathbf{C}'(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})\widetilde{\mathbf{\Sigma}}^{-1}(\widetilde{\mathbf{B}} - \widetilde{\mathbf{B}}_G\mathbf{G})'\mathbf{C}\right]^{-1}(\mathbf{C}'\widetilde{\mathbf{B}}_G\mathbf{M} - \mathbf{\Delta})/2.$$

Outline of Proof: First verify that the LRT does not depend on the marginal distribution of $\mathbf{Z}_2$. *That is, the test depends only on the distribution of* $\mathbf{Z}_1$ *conditional on* $\mathbf{Z}_2$. *Second, write* $\mathbf{M}$ *as* $\mathbf{M} = \mathbf{U}\mathbf{F}$ *for some* $\mathbf{F}$. *Recall that* $\mathbf{M} \in \mathcal{R}(\mathbf{G}) = \mathcal{R}(\mathbf{U})$. *The conditional model can be written as*

$$\mathbf{Z}_1 = \begin{pmatrix}\mathbf{X} & \mathbf{Z}_2\end{pmatrix}\begin{pmatrix}\mathbf{B}\mathbf{U} \\ \mathbf{\Gamma}\end{pmatrix} + \mathbf{W}_{1\cdot 2}^{*},$$

where $\text{disp}(\mathbf{W}_{1\cdot 2}^{*}) = \mathbf{\Omega}_{11\cdot 2} \otimes \mathbf{I}_n$. *Finally, use Theorems 8.9 and 8.10 to construct the LRT of*

$$\text{H}_0\text{:}\ \begin{pmatrix}\mathbf{C}' & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{B}\mathbf{U} \\ \mathbf{\Gamma}\end{pmatrix}\mathbf{F} = \mathbf{\Delta} \text{ vs. } \text{H}_a\text{:}\ \begin{pmatrix}\mathbf{C}' & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{B}\mathbf{U} \\ \mathbf{\Gamma}\end{pmatrix}\mathbf{F} \neq \mathbf{\Delta}.$$

□

8.9 GENERALIZED ANALYSES OF LONGITUDINAL DATA

8.9.1 Introduction to Proc Mixed

Chapter 9

SELECTED INFERENCE ON COVARIANCE MATRICES

9.1 LR TESTS FOR SELECTED COVARIANCE STRUCTURES

Tests for the following hypotheses will be derived in class.

1. Tests for a specified structure: $\mathbf{\Sigma} = \mathbf{\Sigma}_0$.
2. Tests for sphericity: $\mathbf{\Sigma} = \sigma^2\mathbf{I}_p$ or $\mathbf{C}'\mathbf{\Sigma}\mathbf{C} = \sigma^2\mathbf{I}_{p-1}$.
3. Tests for intraclass structure: $\mathbf{\Sigma} = \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{J}]$. Hints:

$$\begin{aligned}
\mathbf{\Sigma} &= \sigma^2(1-\rho)\mathbf{I}_p + \sigma^2\rho\mathbf{J} \Longleftrightarrow \\
\mathbf{\Sigma}^{-1} &= \frac{1}{\sigma^2(1-\rho)}\mathbf{I}_p - \frac{\rho}{\sigma^2(1-\rho)[1+(p-1)\rho]}\mathbf{J} \\
&= \frac{1}{\theta_2}\mathbf{I}_p - \frac{(\theta_1-\theta_2)}{p\theta_1\theta_2}\mathbf{J} \text{ where } \theta_1 = \sigma^2[1+(p-1)\rho] \text{ and } \theta_2 = \sigma^2(1-\rho) \\
|\mathbf{\Sigma}| &= \sigma^2[1+(p-1)\rho][(1-\rho)\sigma^2]^{p-1} = \theta_1\theta_2^{p-1}.
\end{aligned}$$

4. Testing equality of covariance matrices: $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 \cdots = \mathbf{\Sigma}_k$.

9.2 CANONICAL CORRELATION AND BLOCKWISE INDEPENDENCE

9.2.1 Review of Bivariate Correlation

Suppose that

$$\text{vec}(\mathbf{Y}) \sim \text{N}\left[\text{vec}(\mathbf{XB}), \mathbf{\Sigma} \otimes \mathbf{I}_n\right],$$

where $\mathbf{Y}$ is $n \times 2$, $\mathbf{X}$ is $n \times p$ with rank r_x, and $\mathbf{\Sigma}$ is 2×2. We are interested in testing the hypothesis that the two columns of $\mathbf{Y}$ are independent. Under normality, the columns are independent iff $\mathbf{\Sigma}$ is a diagonal matrix. Write $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \{\sigma_{ij}\}$. We want to test $\text{H}_0\colon \sigma_{12} = 0$.

Theorem 9.1 *The LR test of* $\text{H}_0\colon \sigma_{12} = 0$ *is the following: reject* H_0 *for small*

$$\Lambda = (1-r^2)^{\frac{n}{2}},$$

where r is the sample correlation coefficient:

$$r = \frac{s_{12}}{(s_{11}s_{22})^{\frac{1}{2}}},$$

and $\mathbf{S} = \{s_{ij}\}$ is the usual unbiased estimator of $\mathbf{\Sigma}$. An equivalent test is to reject H_0 *for large*

$$F = \frac{(n - r_x - 1)r^2}{1 - r^2}.$$

□

The statistic r is called the sample Pearson product moment correlation coefficient. The population coefficient is usually denoted by ρ:

$$\rho = \frac{\sigma_{12}}{(\sigma_{11}\sigma_{22})^{\frac{1}{2}}}.$$

Using the Cauchy-Schwartz inequality, it is easy to show that $\rho \in [-1, 1]$.

9.2.2 Review of Multiple Correlation

Suppose that

$$\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}\left[\mathrm{vec}(\mathbf{XB}), \mathbf{\Sigma} \otimes \mathbf{I}_n\right],$$

where $\mathbf{Y}$ is $n \times d$, $\mathbf{X}$ is $n \times p$ with rank r_x, and $\mathbf{\Sigma}$ is $d \times d$. We are interested in testing the hypothesis that the first column of $\mathbf{Y}$ is independent of the remaining columns. Partition $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

where σ_{11} is a scalar. Partition $\mathbf{Y}$ and $\mathbf{B}$ accordingly: $\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{Y}_2 \end{pmatrix}$, where $\mathbf{y}_1$ is an $n \times 1$ vector and $\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1 & \mathbf{B}_2 \end{pmatrix}$, where $\boldsymbol{\beta}_1$ is a p-vector. Under normality, the columns are independent iff $\mathbf{\Sigma}_{12} = \mathbf{0}$.

Theorem 9.2 *The LR test of* $\mathrm{H}_0 \colon \mathbf{\Sigma}_{12} = \mathbf{0}$ *is the following: reject* H_0 *for small*

$$\Lambda = (1 - R^2)^{\frac{n}{2}},$$

where R^2 is the sample squared multiple correlation coefficient:

$$R^2 = \frac{\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}}{s_{11}},$$

$$\mathbf{S} = \begin{pmatrix} s_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}}{n - r_x},$$

and $\mathbf{H} = \mathrm{ppo}(\mathbf{X})$. An equivalent test is to reject H_0 *for large*

$$F = \frac{(n - r_x - d + 1)R^2}{(d - 1)(1 - R^2)}.$$

□

Theorem 9.3 *The statistic R^2 is the maximum squared sample Pearson product moment correlation coefficient between the first column of $\mathbf{Y}$ and an arbitrary linear combination of the remaining columns. Let $r^2(\mathbf{t})$ be the squared sample correlation between $\mathbf{y}_1$ and $\mathbf{Y}_2\mathbf{t}$, where $\mathbf{t}$ is a $(d-1) \times 1$ vector. Then,*

$$R^2 = \max_{\mathbf{t} \neq \mathbf{0}} r^2(\mathbf{t}).$$

The squared population coefficient is usually denoted by ρ^2:

$$\rho^2 = \frac{\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}}{\sigma_{11}}.$$

□

Theorem 9.4 *Conditional on* $\mathbf{Y}_2$*,* F *in Theorem 9.2 is distributed as*

$$F|\mathbf{Y}_2 \sim F_{d-1,n-d,\lambda},$$

where

$$\lambda = \frac{\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}}{2\sigma_{11}(1-\rho^2)},$$

and $\mathbf{H} = \text{ppo}(\mathbf{X})$*.*

Sketch of Proof: Let $\mathbf{H}^* = \text{ppo}(\mathbf{X}^*)$ *where* $\mathbf{X}^* = \begin{pmatrix}\mathbf{X} & \mathbf{Y}_2\end{pmatrix}$*. It can be shown that* $\mathbf{H}^* = \mathbf{H} + \mathbf{H}_{2\cdot1}$*, where* $\mathbf{H}_{2\cdot1} = \text{ppo}\left[(\mathbf{I}_n - \mathbf{H})\mathbf{Y}_2\right]$*. The test statistic can be written as*

$$F = \frac{(n - r_x - d + 1)\mathbf{y}_1'\mathbf{H}_{2\cdot1}\mathbf{y}_1}{(d-1)\mathbf{y}_1'(\mathbf{I}_n - \mathbf{H}^*)\mathbf{y}_1}.$$

It can be shown that the conditional distribution of $\mathbf{y}_1$ *given* $\mathbf{Y}_2$ *is*

$$\mathbf{y}_1|\mathbf{Y}_2 \sim \text{N}(\mathbf{X}^*\boldsymbol{\beta}^*, \sigma^2\mathbf{I}_n),$$

where

$$\boldsymbol{\beta}^* = \begin{pmatrix}\boldsymbol{\beta}_1 - \mathbf{B}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\end{pmatrix},$$

and $\sigma^2 = \sigma_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$*. It is readily established that the numerator and denominator of* F *are conditionally independent and distributed as chi squared random variables.*

□

Corollary 1: The expectation of the conditional noncentrality parameter is

$$\text{E}(\lambda) = \frac{(n - r_x)\rho^2}{2(1-\rho^2)}.$$

Corollary 2: If H_0 is true, then $F \sim F_{d-1,n-r_x-d+1,0}$, unconditionally.

9.2.3 Blockwise Independence: Two Blocks

The above procedure can be generalized as follows. Consider the model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where $\mathbf{Y}$ is $n \times p + q$, $\mathbf{X}$ is $n \times h$ with rank-r, and $\text{vec}(\mathbf{U}) \sim \text{N}\left[\mathbf{0}, (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)\right]$. Partition $\mathbf{Y}$ and $\boldsymbol{\Sigma}$ as $\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{pmatrix}$ and

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

where $\mathbf{Y}_1$ is $n \times p$, $\mathbf{Y}_2$ is $n \times q$, $\boldsymbol{\Sigma}_{11}$ is $p \times p$, and $\boldsymbol{\Sigma}_{22}$ is $q \times q$. We wish to test the hypothesis that $\mathbf{Y}_1$ is independent of $\mathbf{Y}_2$. Under normality, the null hypothesis can be written as $\text{H}_0\colon \boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Theorem 9.5 *The LR test of* $\text{H}_0\colon \boldsymbol{\Sigma}_{12} = \mathbf{0}$ *is the following: reject* H_0 *for small*

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

where

$$\mathbf{A} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y},$$

$\mathbf{H}_x = \text{ppo}(\mathbf{X})$*, and* $\mathbf{A}$ *is partitioned identically to* $\boldsymbol{\Sigma}$*.*

□

Theorem 9.6 *Under* H_0*,* U *is distributed as*

$$U \sim \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|},$$

where $\mathbf{E} \sim \mathrm{W}_q(n-r-p, \mathbf{I}_q)$*,* $\mathbf{H} \sim \mathrm{W}_q(p, \mathbf{I}_q)$*, and* $\mathbf{E} \perp\!\!\!\perp \mathbf{H}$*. It follows that* $U \sim U(q, p, n-r-p)$*.*

Proof: First, note that

$$U = \frac{|\mathbf{A}_{11}| \times |\mathbf{A}_{22\cdot 1}|}{|\mathbf{A}_{11}| \times |\mathbf{A}_{22}|} = \frac{|\mathbf{A}_{22\cdot 1}|}{|\mathbf{A}_{22}|} = \frac{|\mathbf{A}_{22\cdot 1}|}{|\mathbf{A}_{22\cdot 1} + \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|}.$$

Now use Corollary 2 to Theorem 4.9 to show that

1. $\mathbf{A}_{22\cdot 1} \sim \mathrm{W}_q(n-r-p, \mathbf{\Sigma}_{22\cdot 1})$ *and*

2. $\mathbf{A}_{22\cdot 1} \perp\!\!\!\perp \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$.

Note that $\mathbf{I} - \mathbf{H}_x$ *can be factored as*

$$\mathbf{I} - \mathbf{H}_x = \mathbf{V}\mathbf{V}', \text{ where } \mathbf{V}\colon n \times (n-r) \text{ satisfies } \mathbf{V}'\mathbf{V} = \mathbf{I}_{n-r}.$$

Accordingly, $\mathbf{A}$ *can be written as*

$$\mathbf{A} = \mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y} = \mathbf{Z}'\mathbf{Z}, \text{ where } \mathbf{Z} = \mathbf{V}'\mathbf{Y} \text{ and } \mathrm{vec}(\mathbf{Z}) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}_{n-r}).$$

Partition $\mathbf{Z}$ *as* $\mathbf{Z} = \begin{pmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{pmatrix}$*, where* $\mathbf{Z}_1$ *is* $(n-r) \times p$ *and* $\mathbf{Z}_2$ *is* $(n-r) \times q$*. The distribution of* $\mathbf{Z}_2$ *conditional on* $\mathbf{Z}_1$ *is*

$$\mathrm{vec}(\mathbf{Z}_2)|\mathbf{Z}_1 = \ddot{\mathbf{Z}}_1 \sim \mathrm{N}\left[\mathrm{vec}\left(\ddot{\mathbf{Z}}_1\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right), \mathbf{\Sigma}_{22\cdot 1} \otimes \mathbf{I}_{n-r}\right].$$

It follows that

$$\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \mathbf{Z}_2' \,\mathrm{ppo}(\mathbf{Z}_1)\mathbf{Z}_2 \text{ and}$$

$$\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|\mathbf{Z}_1 = \ddot{\mathbf{Z}}_1 \sim \mathrm{W}_q(p, \mathbf{\Sigma}_{22\cdot 1}, \mathbf{\Lambda}), \text{ where}$$

$$\mathbf{\Lambda} = \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\ddot{\mathbf{A}}_{11}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22\cdot 1}^{-1}.$$

If H_0 *is true, then* $\mathbf{\Sigma}_{12} = \mathbf{0}$ *and*

$$\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \sim \mathrm{W}_q(p, \mathbf{\Sigma}_{22\cdot 1})$$

unconditionally.

□

Example

Below is a data set concerning various physiological and fitness measures. The data were obtained by A.C. Linnerud on 20 middle aged men, all belonging to a fitness club.

Linnerud (1985) Data

Weight	Waist	Pulse	Chinups	Situps	Jumps
191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43

We are interested in determining the relationships between the three physiological variables and the three performance variables. The model is

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\mu}' + \mathbf{U}.$$

The error matrix $\mathbf{A} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$ is

$$\mathbf{A} = 10^4 \times \begin{pmatrix} 1.1583 & 0.1307 & -0.1237 & -0.0966 & -1.4473 & -0.5444 \\ 0.1307 & 0.0195 & -0.0155 & -0.0178 & -0.2457 & -0.0597 \\ -0.1237 & -0.0155 & 0.0988 & 0.0109 & 0.1929 & 0.0245 \\ -0.0966 & -0.0178 & 0.0109 & 0.0531 & 0.4372 & 0.2553 \\ -1.4473 & -0.2457 & 0.1929 & 0.4372 & 7.4377 & 4.0793 \\ -0.5444 & -0.0597 & 0.0245 & 0.2553 & 4.0793 & 4.9958. \end{pmatrix}$$

Using Matlab $|\mathbf{A}| = 9.1452 \times 10^{19}$, $|\mathbf{A}_{11}| = 4.6585 \times 10^8$, and $|\mathbf{A}_{22}| = 5.6027 \times 10^{11}$. Thus $U = 0.350391$.

Under $\mathrm{H}_0\colon \boldsymbol{\Sigma}_{12} = \mathbf{0}$, U is distributed as $U \sim U(3, 3, 16)$. The parameters of Rao's F approximation (see Rencher, 2002, equation 6.15) are $t = 2.4337$, $w = 15.5$, $df_1 = 9$, and $df_2 = 34.2229$. Thus, the observed F statistic is $F = 2.0482$. This statistic can be compared to the F distribution with 9 and 34.2229 degrees of freedom. The p-value is, approximately, .06324. Hence, at $\alpha = .05$, the null cannot be rejected.

9.2.4 Blockwise Independence: k Blocks

Generalization from 2 to k blocks is straightforward. The above procedure can be generalized as follows. Consider the model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where $\mathbf{Y}$ is $n \times d$, $\mathbf{X}$ is $n \times p$ with rank-r, and $\mathrm{vec}(\mathbf{U}) \sim \mathrm{N}\left[\mathbf{0}, (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)\right]$. Partition $\mathbf{Y}$ and $\boldsymbol{\Sigma}$ as $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_k \end{pmatrix}$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{k1} & \boldsymbol{\Sigma}_{k2} & \cdots & \boldsymbol{\Sigma}_{kk} \end{pmatrix},$$

where $\mathbf{Y}_j$ is $n \times d_j$, $\boldsymbol{\Sigma}_{ij}$ is $d_i \times d_j$, and $\sum_{j=1}^k d_j = d$. We wish to test the hypothesis that $\mathbf{Y}_i$ is independent of $\mathbf{Y}_j$ for all $i \neq j$. Under normality, the null hypothesis can be written as

$$\mathrm{H}_0\colon \boldsymbol{\Sigma} = \bigoplus_{j=1}^k \boldsymbol{\Sigma}_{jj}.$$

Theorem 9.7 *The LR test of* H_0 *is the following: reject* H_0 *for small*

$$U = \frac{|\mathbf{A}|}{\prod_{j=1}^{k} |\mathbf{A}_{jj}|},$$

where

$$\mathbf{A} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y},$$

$\mathbf{H}_x = \mathrm{ppo}(\mathbf{X})$*, and* $\mathbf{A}$ *is partitioned identically to* $\mathbf{\Sigma}$*.*
Proof: HW.

Multiplying the numerator and denominator of U by the same constants reveals that U also can be written as

$$U = \frac{|\mathbf{S}|}{\prod_{j=1}^{k} |\mathbf{S}_{jj}|} = \frac{|\widehat{\mathbf{R}}|}{\prod_{j=1}^{k} |\widehat{\mathbf{R}}_{jj}|},$$

where $\mathbf{S}$ is the sample covariance matrix, $\mathbf{S} = \mathbf{A}/(n-r)$, and $\widehat{\mathbf{R}}$ is the sample correlation matrix.

The Bartlett-corrected test is to reject H_0 if

$$-(n-r)c\ln(U) > \chi^2_{1-\alpha,f}, \text{ where } f = \frac{1}{2}\left(d^2 - \sum_{j=1}^{k} d_j^2\right),$$

$$c = 1 - \frac{1}{12f(n-r)}(4f + 3g), \text{ and } g = d^3 - \sum_{j=1}^{k} d_j^3.$$

9.2.5 Canonical Correlation

Again consider the model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where $\mathbf{Y}$ is $n \times p+q$, $\mathbf{X}$ is $n \times h$ with rank-r, and $\mathrm{vec}(\mathbf{U}) \sim \mathrm{N}[\mathbf{0}, (\mathbf{\Sigma} \otimes \mathbf{I}_n)]$. Partition $\mathbf{Y}$ and $\mathbf{\Sigma}$ as $\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_1 & \mathbf{Y}_2\end{pmatrix}$ and

$$\mathbf{\Sigma} = \begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{pmatrix},$$

where $\mathbf{Y}_1$ is $n \times p$, $\mathbf{Y}_2$ is $n \times q$, $\mathbf{\Sigma}_{11}$ is $p \times p$, and $\mathbf{\Sigma}_{22}$ is $q \times q$. The LR test for blockwise independence is one way to test the hypothesis that $\mathbf{Y}_1$ is independent of $\mathbf{Y}_2$.

An alternative way of testing that $\mathbf{Y}_1$ is independent of $\mathbf{Y}_2$ is through canonical correlations. Consider a linear combination of the responses in $\mathbf{Y}_1$, say $\mathbf{z}_1 = \mathbf{Y}_1\boldsymbol{\ell}_1$ where $\boldsymbol{\ell}_1$ is $p \times 1$. Consider a linear combination of the responses in $\mathbf{Y}_2$, say $\mathbf{z}_2 = \mathbf{Y}_2\boldsymbol{\ell}_2$ where $\boldsymbol{\ell}_2$ is $q \times 1$. Let $\mathbf{Z} = \begin{pmatrix}\mathbf{z}_1 & \mathbf{z}_2\end{pmatrix}$. Let

$$\mathbf{L} = \boldsymbol{\ell}_1 \oplus \boldsymbol{\ell}_2 = \begin{pmatrix}\boldsymbol{\ell}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\ell}_2\end{pmatrix}.$$

Then, $\mathbf{Z} = \mathbf{YL}$ and

$$\mathrm{disp}(\mathbf{Z}) = (\mathbf{L}' \otimes \mathbf{I}_n)\,\mathrm{disp}(\mathbf{Y})(\mathbf{L} \otimes \mathbf{I}_n) = (\mathbf{L}'\mathbf{\Sigma}\mathbf{L} \otimes \mathbf{I}_n) = (\mathbf{\Omega} \otimes \mathbf{I}_n),$$

where

$$\mathbf{\Omega} = \{\omega_{ij}\} = \mathbf{L}'\mathbf{\Sigma}\mathbf{L} = \begin{pmatrix}\boldsymbol{\ell}_1'\mathbf{\Sigma}_{11}\boldsymbol{\ell}_1 & \boldsymbol{\ell}_1'\mathbf{\Sigma}_{12}\boldsymbol{\ell}_2 \\ \boldsymbol{\ell}_2'\mathbf{\Sigma}_{21}\boldsymbol{\ell}_1 & \boldsymbol{\ell}_2'\mathbf{\Sigma}_{22}\boldsymbol{\ell}_2\end{pmatrix}.$$

If $\mathbf{Y}_1$ is independent of $\mathbf{Y}_2$, then $\omega_{12} = 0$ regardless of how we choose $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$. If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are not independent, then $\omega_{12} \neq 0$ for some choice of $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$. Thus, the null and alternative hypotheses can be written as

$$\mathrm{H}_0 \colon \mathbf{\Sigma}_{12} = \mathbf{0} \iff \boldsymbol{\ell}_1'\mathbf{\Sigma}_{12}\boldsymbol{\ell}_2 = 0 \;\forall\; \boldsymbol{\ell}_1, \boldsymbol{\ell}_2$$

and

$$\mathrm{H}_a\colon \boldsymbol{\Sigma}_{12} \neq \mathbf{0} \Longleftrightarrow \boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_2 \neq 0 \text{ for some } \boldsymbol{\ell}_1, \boldsymbol{\ell}_2.$$

These hypotheses are in the union intersection form:

$$\mathrm{H}_0\colon \bigcap_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} \boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_2 = 0 \text{ versus } \mathrm{H}_a\colon \bigcup_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} \boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_2 \neq 0.$$

Accordingly, a union-intersection test can be constructed as follows:

1. Construct a test of $\mathrm{H}_0\colon \boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_2 = 0$ versus $\mathrm{H}_a\colon \boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_2 \neq 0$ for an a priori pair $\boldsymbol{\ell}_1$, $\boldsymbol{\ell}_2$. Denote the test statistic by $V_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}$ where H_0 is rejected for large V.
2. The UI test is to reject $\mathrm{H}_0\colon \boldsymbol{\Sigma}_{12} = \mathbf{0}$ for large
$$V = \max_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} V_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}.$$
3. Find a convenient way to compute V and find the null distribution of V.

For an a priori $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$, the problem reduces to testing independence between two columns. From prior work, we know that the likelihood ratio test is to reject $\mathrm{H}_0\colon \omega_{12} = 0$ for large $V_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} = r^2_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}$ where $r_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}$ is the sample correlation between $\mathbf{Y}_1\boldsymbol{\ell}_1$ and $\mathbf{Y}_2\boldsymbol{\ell}_2$:

$$V_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} = r^2_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} = \frac{\left(\boldsymbol{\ell}_1'\mathbf{S}_{12}\boldsymbol{\ell}_2\right)^2}{(\boldsymbol{\ell}_1'\mathbf{S}_{11}\boldsymbol{\ell}_1)(\boldsymbol{\ell}_2'\mathbf{S}_{22}\boldsymbol{\ell}_2)},$$

where $\mathbf{S}$ is the sample covariance matrix. Thus, the UI test statistic is

$$V = \max_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} \frac{\left(\boldsymbol{\ell}_1'\mathbf{S}_{12}\boldsymbol{\ell}_2\right)^2}{(\boldsymbol{\ell}_1'\mathbf{S}_{11}\boldsymbol{\ell}_1)(\boldsymbol{\ell}_2'\mathbf{S}_{22}\boldsymbol{\ell}_2)} = \mathrm{ch}_1(\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}),$$

where $\mathrm{ch}_1(\cdot)$ is the maximum characteristic root function. The statistic V is called the first squared sample canonical correlation. It is the maximum squared correlation between a linear combination of the columns of $\mathbf{Y}_1$ and a linear combination of the columns of $\mathbf{Y}_2$. If either $p = 1$ or $q = 1$, then V reduces to a squared multiple correlation. If $p = 1$ and $q = 1$, then V reduces to a squared simple correlation.

Theorem 9.8 *Conditional on $\mathbf{Y}_1$, V has the same distribution as the maximum root of $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$ where $\mathbf{E}$ and $\mathbf{H}$ are independently distributed as*

$$\mathbf{E} \sim \mathrm{W}_q(n-r-p, \boldsymbol{\Sigma}_{22\cdot 1}) \quad \text{and} \quad \mathbf{H}|\mathbf{A}_{11} \sim \mathrm{W}_q(p, \boldsymbol{\Sigma}_{22\cdot 1}, \boldsymbol{\Sigma}_{22\cdot 1}^{-1}\boldsymbol{\Lambda}),$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{A}_{11}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

□

Corollary: Under H_0, V has the same distribution as the maximum root of $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$ where $\mathbf{E}$ and $\mathbf{H}$ are independently distributed as

$$\mathbf{E} \sim \mathrm{W}_q(n-r-p, \mathbf{I}) \quad \text{and} \quad \mathbf{H} \sim \mathrm{W}_q(p, \mathbf{I}).$$

Population canonical correlations are obtained by substituting $\boldsymbol{\Sigma}$ for $\mathbf{S}$. Let

$$\rho_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2} = \frac{\omega_{12}}{(\omega_{11}\omega_{22})^{\frac{1}{2}}}.$$

That is, $\rho_{\boldsymbol{\ell}_1,\boldsymbol{\ell}_2}$ is the correlation between the two columns of $\mathbf{Z}$. One way to summarize the relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is to find the linear combinations, $\mathbf{z}_1 = \mathbf{Y}_1\boldsymbol{\ell}_1$ and $\mathbf{z}_2 = \mathbf{Y}_2\boldsymbol{\ell}_2$, which have the largest correlation. The linear combinations, $\mathbf{z}_1$ and $\mathbf{z}_2$ are called the first canonical variates and their correlation is called the first canonical correlation. It may be that the relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is not adequately summarized by a single correlation. In which case, we can construct additional canonical variates and correlations.

Denote the rank of $\boldsymbol{\Sigma}_{12}$ by s and denote the ordered characteristic roots and vectors of $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ by $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_s)$ and $\mathbf{L}_1 = \begin{pmatrix}\boldsymbol{\ell}_{11} & \cdots & \boldsymbol{\ell}_{1m}\end{pmatrix}$, where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_s$. That is,

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1 = \mathbf{L}_1\boldsymbol{\Lambda}.$$

In a similar manner, denote the characteristic vectors of $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ by $\mathbf{L}_2 = \begin{pmatrix}\boldsymbol{\ell}_{21} & \cdots & \boldsymbol{\ell}_{2m}\end{pmatrix}$:

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{L}_2 = \mathbf{L}_2\boldsymbol{\Lambda}.$$

Note that the characteristic roots corresponding to $\mathbf{L}_1$ are identical to those corresponding to $\mathbf{L}_2$. The following results can be established:

1. If the eigenvalues in $\boldsymbol{\Lambda}$ are distinct, then $\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1$ and $\mathbf{L}_2'\boldsymbol{\Sigma}_{22}\mathbf{L}_2$ are each diagonal. If the eigenvalues are not distinct, then the eigenvectors $\mathbf{L}_1$ and $\mathbf{L}_2$ can be chosen such that $\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1$ and $\mathbf{L}_2'\boldsymbol{\Sigma}_{22}\mathbf{L}_2$ are each diagonal. To verify this claim, define $\boldsymbol{\Xi}$ as $\boldsymbol{\Xi} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Then,

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1 = \mathbf{L}_1\boldsymbol{\Lambda}$$

$$\Longrightarrow \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Xi}\boldsymbol{\Xi}'\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 = \mathbf{L}_1\boldsymbol{\Lambda}$$

$$\Longrightarrow \boldsymbol{\Xi}\boldsymbol{\Xi}'\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1\boldsymbol{\Lambda}$$

$$\Longrightarrow \text{ the columns of } \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 \text{ are eigenvectors of } \boldsymbol{\Xi}\boldsymbol{\Xi}'.$$

   Note that $\boldsymbol{\Xi}\boldsymbol{\Xi}'$ is symmetric. Accordingly, if the eigenvalues λ_i for $i = 1, \ldots, s$ are distinct, then columns of $\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1$ are orthogonal. Otherwise, the columns of $\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1$ can be chosen to be orthogonal. Therefore, $\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1$ is a diagonal matrix.

2. One way to compute the eigenvectors $\mathbf{L}_1$ and $\mathbf{L}_2$ is as follows. Write the full-rank SVD of $\boldsymbol{\Xi}$ as

$$\boldsymbol{\Xi} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} = \mathbf{W}_1\boldsymbol{\Theta}\mathbf{W}_2',$$

   where $\mathbf{W}_1$ is $p \times s$, $\mathbf{W}_2$ is $q \times s$, and $\mathbf{W}_1'\mathbf{W}_1 = \mathbf{W}_2'\mathbf{W}_2 = \mathbf{I}_s$. Then

$$\boldsymbol{\Xi}\boldsymbol{\Xi}'\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1\boldsymbol{\Lambda}$$

$$\Longrightarrow \mathbf{W}_1\boldsymbol{\Theta}^2\mathbf{W}_1'\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1\boldsymbol{\Lambda}$$

$$\Longrightarrow \text{the columns of } \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 \text{ are the eigenvectors of } \mathbf{W}_1\boldsymbol{\Theta}^2\mathbf{W}_1' \text{ and } \boldsymbol{\Lambda} = \boldsymbol{\Theta}^2$$

$$\Longrightarrow \text{ the columns of } \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1 \text{ can be chosen to be proportional to the columns of } \mathbf{W}_1.$$

   It follows that $\mathbf{L}_1$ can be chosen to satisfy

$$\boldsymbol{\Sigma}_{11}^{1/2}\mathbf{L}_1\mathbf{D}_1 = \mathbf{W}_1$$

   for some $s \times s$ diagonal matrix, $\mathbf{D}_1$ Accordingly,

$$\mathbf{L}_i = \boldsymbol{\Sigma}_{ii}^{-\frac{1}{2}}\mathbf{W}_i\mathbf{D}_i^{-1}$$

   for $i = 1, 2$, where $\mathbf{D}_i$ is an arbitrary matrix.

3. Partition the matrix of regression coefficients as $\mathbf{B} = (\mathbf{B}_1\ \mathbf{B}_2)$, where $\mathbf{B}_1$ is $h \times p$ and $\mathbf{B}_2$ is $h \times q$. Define $\mathbf{Z}$ by

$$\mathbf{Z} \stackrel{\text{def}}{=} (\mathbf{Y} - \mathbf{X}\mathbf{B})\begin{pmatrix}\mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2\end{pmatrix} = \begin{pmatrix}(\mathbf{Y}_1 - \mathbf{X}\mathbf{B}_1) & (\mathbf{Y}_2 - \mathbf{X}\mathbf{B}_2)\end{pmatrix}\begin{pmatrix}\mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2\end{pmatrix}$$

$$= \begin{pmatrix}(\mathbf{Y}_1 - \mathbf{X}\mathbf{B}_1)\mathbf{L}_1 & (\mathbf{Y}_2 - \mathbf{X}\mathbf{B}_2)\mathbf{L}_2\end{pmatrix} = \begin{pmatrix}\mathbf{Z}_1 & \mathbf{Z}_2\end{pmatrix}.$$

The $n \times s$ matrix $\mathbf{Z}$ contains the centered canonical variates. Note that

$$\text{Disp}(\mathbf{Z}) = \begin{pmatrix} \mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1 & \mathbf{L}_1'\boldsymbol{\Sigma}_{12}\mathbf{L}_2 \\ \mathbf{L}_2'\boldsymbol{\Sigma}_{21}\mathbf{L}_1 & \mathbf{L}_2'\boldsymbol{\Sigma}_{22}\mathbf{L}_2 \end{pmatrix} \otimes \mathbf{I}_n$$

$$= \begin{pmatrix} \mathbf{D}_1^{-1}\mathbf{W}_1'\boldsymbol{\Sigma}_{11}^{-\frac{1}{1}}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{W}_1\mathbf{D}_1^{-1} & \mathbf{D}_1^{-1}\mathbf{W}_1'\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{W}_2\mathbf{D}_2^{-1} \\ \mathbf{D}_2^{-1}\mathbf{W}_2'\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{W}_1\mathbf{D}_1^{-1} & \mathbf{D}_1^{-2}\mathbf{W}_2'\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{W}_2\mathbf{D}_2^{-1} \end{pmatrix} \otimes \mathbf{I}_n$$

$$= \begin{pmatrix} \mathbf{D}_1^{-2} & \mathbf{D}_1^{-1}\boldsymbol{\Theta}\mathbf{D}_2^{-1} \\ \mathbf{D}_2^{-1}\boldsymbol{\Theta}\mathbf{D}_1^{-1} & \mathbf{D}_2^{-2} \end{pmatrix} \otimes \mathbf{I}_n.$$

If $\mathbf{D}_i$ is chosen to be $\mathbf{I}_s$, then

$$\text{Disp}(\mathbf{Z}) = \begin{pmatrix} \mathbf{I}_s & \boldsymbol{\Theta} \\ \boldsymbol{\Theta} & \mathbf{I}_s \end{pmatrix} \otimes \mathbf{I}_n.$$

4. Choosing $\mathbf{D}_i$ to be an identity is equivalent to solving the eigen-equations

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1 = \mathbf{L}_1\boldsymbol{\Lambda} \text{ and } \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{L}_2 = \mathbf{L}_2\boldsymbol{\Lambda}$$

and then replacing $\mathbf{L}_1$ by

$$\mathbf{L}_1^* = \mathbf{L}_1(\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1)^{-\frac{1}{2}}$$

and replacing $\mathbf{L}_2$ by

$$\mathbf{L}_2^* = \mathbf{L}_2(\mathbf{L}_2'\boldsymbol{\Sigma}_{22}\mathbf{L}_2)^{-\frac{1}{2}}.$$

Note that

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1 = \mathbf{L}_1\boldsymbol{\Lambda}$$

$$\implies \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1\left(\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1\right)^{-\frac{1}{2}} = \mathbf{L}_1\boldsymbol{\Lambda}\left(\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1\right)^{-\frac{1}{2}}$$

$$\implies \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1\left(\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1\right)^{-\frac{1}{2}} = \mathbf{L}_1\left(\mathbf{L}_1'\boldsymbol{\Sigma}_{11}\mathbf{L}_1\right)^{-\frac{1}{2}}\boldsymbol{\Lambda}$$

$$\implies \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1^* = \mathbf{L}_1^*\boldsymbol{\Lambda}$$

because diagonal matrices commute. Also, $\mathbf{L}_1^{*\prime}\boldsymbol{\Sigma}_{11}\mathbf{L}_1^* = \mathbf{L}_2^{*\prime}\boldsymbol{\Sigma}_{22}\mathbf{L}_2^* = \mathbf{I}_s$. For convenience, $\mathbf{L}_1^*$ and $\mathbf{L}_2^*$ will be denoted simply as $\mathbf{L}_1$ and $\mathbf{L}_2$.

5. $\mathbf{L}_1'\boldsymbol{\Sigma}_{12}\mathbf{L}_2 = \boldsymbol{\Theta} = \boldsymbol{\Lambda}^{\frac{1}{2}}$. Thus, λ_i is the squared correlation between $\mathbf{Y}_1\boldsymbol{\ell}_{1i}$ and $\mathbf{Y}_2\boldsymbol{\ell}_{2i}$, and $\mathbf{Y}_1\boldsymbol{\ell}_{1i} \perp\!\!\!\perp \mathbf{Y}_2\mathbf{w}_{2i'}$ for $i \neq i'$.

6. $\mathbf{L}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1 = \mathbf{L}_2'\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{L}_2 = \boldsymbol{\Lambda}$.

7. $\mathbf{L}_2 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{L}_1\boldsymbol{\Lambda}^{-\frac{1}{2}}$ and $\mathbf{L}_1 = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\mathbf{L}_2\boldsymbol{\Lambda}^{-\frac{1}{2}}$.

8. Denote the $n \times (i-1)$ matrix consisting of the first $i-1$ columns of $\mathbf{L}_1$ by $\mathbf{L}_{1i}$. Similarly, denote the $n \times (i-1)$ matrix consisting of the first $i-1$ columns of $\mathbf{L}_2$ by $\mathbf{L}_{2i}$. If $i = 1$, then $\mathbf{L}_{11} = \mathbf{L}_{21} = \mathbf{0}$. Let $V_{1i} = \mathcal{N}(\mathbf{L}_{1i}'\boldsymbol{\Sigma}_{11})$ and let $V_{2i} = \mathcal{N}(\mathbf{L}_{2i}'\boldsymbol{\Sigma}_{22})$. Then,

$$\max_{\boldsymbol{\ell}_1 \in V_{1i}, \boldsymbol{\ell}_2 \in V_{2i}} \frac{\left(\boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_2\right)^2}{(\boldsymbol{\ell}_1'\boldsymbol{\Sigma}_{11}\boldsymbol{\ell}_1)(\boldsymbol{\ell}_2'\boldsymbol{\Sigma}_{22}\boldsymbol{\ell}_2)} = \left(\boldsymbol{\ell}_{1i}'\boldsymbol{\Sigma}_{12}\boldsymbol{\ell}_{2i}\right)^2 = \lambda_i.$$

The i^{th} set of canonical variates is $\mathbf{z}_{1i} = \mathbf{Y}_1\boldsymbol{\ell}_{1i}$ and $\mathbf{z}_{2i} = \mathbf{Y}_2\boldsymbol{\ell}_{2i}$. The squared correlation between $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ is called the i^{th} squared canonical correlation and is equal to λ_i.

9. The nonzero characteristic roots of $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ are identical to the nonzero characteristic roots of $\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$, where $\mathbf{R}$ is the population correlation matrix. The corresponding scaled eigenvectors are called the standardized canonical coefficients.

Example The sample covariance matrix corresponding to the data on page 72 of the notes is

$$\mathbf{S} = 10^3 \times \begin{pmatrix} 0.6096 & 0.0688 & -0.0651 & -0.0509 & -0.7617 & -0.2865 \\ 0.0688 & 0.0103 & -0.0081 & -0.0093 & -0.1293 & -0.0314 \\ -0.0651 & -0.0081 & 0.0520 & 0.0057 & 0.1015 & 0.0129 \\ -0.0509 & -0.0093 & 0.0057 & 0.0279 & 0.2301 & 0.1344 \\ -0.7617 & -0.1293 & 0.1015 & 0.2301 & 3.9146 & 2.1470 \\ -0.2865 & -0.0314 & 0.0129 & 0.1344 & 2.1470 & 2.6294 \end{pmatrix}.$$

Using MATLAB,

▷ S11=S(1:3,1:3);
▷ S22=S(4:6,4:6);
▷ S12=S(1:3,4:6);
▷ S21=S12';
▷ Q1=inv(S11)*S12*inv(S22)*S21;
▷ [W1,Lam]=eig(Q1)

$$W1 = \begin{pmatrix} 0.0635 & -0.2020 & -0.0360 \\ -0.9978 & 0.9757 & 0.7347 \\ 0.0166 & -0.0848 & 0.6775 \end{pmatrix}$$

$$\text{Lam} = \begin{pmatrix} 0.6330 & 0 & 0 \\ 0 & 0.0402 & 0 \\ 0 & 0 & 0.0053 \end{pmatrix}$$

$$\mathbf{L}_1'\mathbf{S}_{11}\mathbf{L}_1 = \begin{pmatrix} 4.0926 & 0 & 0 \\ 0 & 7.0024 & 0 \\ 0 & 0 & 21.6112 \end{pmatrix}$$

$$\mathbf{L}_1(\mathbf{L}_1'\mathbf{S}_{11}\mathbf{L}_1)^{-.5} = \begin{pmatrix} 0.0314 & -0.0763 & -0.0077 \\ -0.4932 & 0.3687 & 0.1580 \\ 0.0082 & -0.0321 & 0.1457 \end{pmatrix}$$

$$\mathbf{L}_1'\mathbf{S}_{12}(\mathbf{S}_{22})^{-1}\mathbf{S}_{21}\mathbf{L}_1 = \begin{pmatrix} 0.6330 & 0 & 0 \\ 0 & 0.0402 & 0 \\ 0 & 0 & 0.0053 \end{pmatrix}$$

$$\mathbf{L}_2 = (\mathbf{S}_{22})^{-1}\mathbf{S}_{21}\mathbf{L}_1\mathbf{\Lambda}^{-.5} = \begin{pmatrix} 0.0661 & -0.0710 & -0.2453 \\ 0.0168 & 0.0020 & 0.0198 \\ -0.0140 & 0.0207 & -0.0082 \end{pmatrix}$$

$$\mathbf{L}_1'\mathbf{S}_{12}\mathbf{L}_2 = \begin{pmatrix} 0.7956 & 0 & 0 \\ 0 & 0.2006 & 0 \\ 0 & 0 & 0.0726 \end{pmatrix}$$

From Table A.10, the 95^{th} percentile of the greatest characteristic root distribution with $s = 3$, $m = -\frac{1}{2}$, and $N = 6$ is difficult to determine. Charts (from Morrison, 1990) will be distributed in class. From chart 11, the critical value is approximately .575. Thus, the first canonical correlation is significant at $\alpha = .05$.

Chapter 10

DISCRIMINANT & CLASSIFICATION ANALYSIS

Consider k populations $P_1, \ldots, P_k$. Let $\mathbf{y}$ be a d-vector, randomly drawn from the i^{th} population. Assume that the density function for $\mathbf{y}$, is $f_i(\mathbf{y})$. The density functions need not be multivariate normal. The densities could be continuous, discrete, or a mixture of continuous and discrete. Suppose that a vector $\mathbf{y}$ is observed, but it is not known from which population the vector was selected. The goal of discriminant analysis is to classify (assign) the vector to the correct population. The following notes are based on Anderson (1958, 1984).

10.1 GENERAL TWO-POPULATION CLASSIFICATION ANALYSIS

10.1.1 Decision Rule, Costs & Risk

Denote the support set for $\mathbf{y}$ by $\mathcal{Y}$. It is assumed that $\mathbf{y}$ has the same support set in both populations. Note that $\mathcal{Y} \in \mathbb{R}^d$. Partition $\mathcal{Y}$ into mutually exclusive and exhaustive subspaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$. The decision rule that we adopt is to classify the observation into P_i if $\mathbf{y} \in \mathcal{Y}_i$. The problem to be solved is how to partition $\mathcal{Y}$. To partition the support set we will use Bayesian methods.

Denote by $C(i|j)$, the cost of classifying an observation into population i when it actually comes from population j. Assume $C(i|j) \geq 0$ for all i, j and $C(i|i) = 0$. The costs can be summarized in a 2×2 table:

		Statisticians Decision	
		P_1	P_2
True	P_1	0	$C(2\|1)$
Population	P_2	$C(1\|2)$	0

Denote by $\Pr(i|j, \mathcal{Y}_1, \mathcal{Y}_2)$, the probability of classifying an observation into population i given that the observation comes from population j using partition $\mathcal{Y} = \mathcal{Y}_1 \bigcup \mathcal{Y}_2$ where, $\mathcal{Y}_1 \bigcap \mathcal{Y}_2 = \emptyset$. These probabilities are called misclassification probabilities. For example, $\Pr(1|1, \mathcal{Y}_1, \mathcal{Y}_2)$ is the probability of correctly classifying an observation from population 1 into population 1 and $\Pr(2|1, \mathcal{Y}_1, \mathcal{Y}_2)$ is the probability of incorrectly classifying an observation from population 1 into population 2.

Let π_i be the probability that a randomly selected observation belongs to population i. The values π_i for $i = 1, 2$ are called prior probabilities and represent the relative sizes of the two populations. Naturally, $\pi_1 + \pi_2 = 1$ must be satisfied.

The expected cost or loss resulting from a classification decision is called risk. The expected loss (risk) conditional on the observation coming from population i is denoted by $r(\mathcal{Y}_1, \mathcal{Y}_2|i)$. Specifically,

$$\begin{aligned} r(\mathcal{Y}_1, \mathcal{Y}_2|1) &= C(1|1) \times \Pr(1|1, \mathcal{Y}_1, \mathcal{Y}_2) + C(2|1) \times \Pr(2|1, \mathcal{Y}_1, \mathcal{Y}_2) \\ &= C(2|1) \times \Pr(2|1, \mathcal{Y}_1, \mathcal{Y}_2), \end{aligned}$$

and

$$\begin{aligned} r(\mathcal{Y}_1, \mathcal{Y}_2|2) &= C(1|2) \times \Pr(1|2, \mathcal{Y}_1, \mathcal{Y}_2) + C(2|2) \times \Pr(2|2, \mathcal{Y}_1, \mathcal{Y}_2) \\ &= C(1|2) \times \Pr(1|2, \mathcal{Y}_1, \mathcal{Y}_2). \end{aligned}$$

The unconditional expected loss (risk) is denoted by $r(\mathcal{Y}_1, \mathcal{Y}_2)$ and is given by

$$r(\mathcal{Y}_1, \mathcal{Y}_2) = r(\mathcal{Y}_1, \mathcal{Y}_2|1) \times \pi_1 + r(\mathcal{Y}_1, \mathcal{Y}_2|2) \times \pi_2.$$

10.1.2 Bayes Procedure

A decision rule is Bayes if its risk is a minimum. To obtain the Bayes rule, write out the risk function and minimize with respect to the partition $\mathcal{Y} = (\mathcal{Y}_1, \mathcal{Y}_2)$. The result is the following.

Theorem 10.1 (Two Population Decision Rule) *The Bayes rule is to classify into P_1 if $\mathbf{y} \in \mathcal{Y}_1$ and to classify into P_2 if $\mathbf{y} \in \mathcal{Y}_2$ where*

$$\mathcal{Y}_1 = \{\mathbf{y}; f_1(\mathbf{y}) \times C(2|1) \times \pi_1 \geq f_2(\mathbf{y}) \times C(1|2) \times \pi_2\},$$

and

$$\mathcal{Y}_2 = \{\mathbf{y}; f_1(\mathbf{y}) \times C(2|1) \times \pi_1 < f_2(\mathbf{y}) \times C(1|2) \times \pi_2\}.$$

If $f_2(\mathbf{y}) \neq 0$, then the regions are

$$\mathcal{Y}_1 = \left\{\mathbf{y}; \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} \geq \frac{C(1|2) \times \pi_2}{C(2|1) \times \pi_1}\right\},$$

and

$$\mathcal{Y}_2 = \left\{\mathbf{y}; \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} < \frac{C(1|2) \times \pi_2}{C(2|1) \times \pi_1}\right\}.$$

Proof in class. □

10.1.3 Admissibility of the Bayes Rule (Optional Section)

In most cases the above Bayes rule is admissible. A rule is admissible if there is no rule that is better. The rule $(\mathcal{Y}_1^*, \mathcal{Y}_2^*)$ is better than $(\mathcal{Y}_1, \mathcal{Y}_2)$ if

1. $r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|i) \leq r(\mathcal{Y}_1, \mathcal{Y}_2|i)\ \forall\ i$, and
2. strict inequality holds for some i.

Note, admissibility is concerned with conditional rather than unconditional risk and, thus, is not dependent on a particular prior. The Bayes rule, by definition, has minimum unconditional risk. However, in some unusual circumstances, it may not have the smallest conditional risk for each i.

If attention is restricted to cases in which $0 < \pi_1 < 1\ \forall\ i$, then the Bayes rule is admissible. Also, if

$$\Pr[f_i(\mathbf{y}) = 0|j] = 0\ \forall\ i \neq j$$

then the Bayes rule is admissible and the class of Bayes rules (indexed by the prior) is minimal complete.

As a counter example, suppose $C(1|2) = 10$, $C(2|1) = 0$, $\pi_1 = 1$, and $\pi_2 = 0$. For any pair of density functions, the Bayes rule is $\mathcal{Y}_1 = \mathcal{Y}$ and $\mathcal{Y}_2 = \emptyset$. That is, the Bayes rule always classifies into P_1. The conditional risks are $r(\mathcal{Y}_1, \mathcal{Y}_2|1) = 0$ and $r(\mathcal{Y}_1, \mathcal{Y}_2|2) = 10$. The unconditional risk is $0 + 10 \times 0 = 0$. Consider $\mathcal{Y}_1^* = \emptyset$ and $\mathcal{Y}_2^* = \mathcal{Y}$. The associated conditional risks are $r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|1) = r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|2) = 0$. Thus, $(\mathcal{Y}_1^*, \mathcal{Y}_2^*)$ is better than $(\mathcal{Y}_1, \mathcal{Y}_2)$.

10.2 TWO NORMAL POPULATIONS

Suppose that

$$\mathbf{y} \sim \begin{cases} \mathrm{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) & \text{if } P_1; \text{ and} \\ \mathrm{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) & \text{if } P_2. \end{cases}$$

Theorem 10.2 (Two Population Linear Discriminant Function) *The Bayes rule is to classify into P_1 if*

$$L(\mathbf{y}) \geq c,$$

where

$$c = \ln\left(\frac{\pi_2 \times C(1|2)}{\pi_1 \times C(2|1)}\right),$$

and

$$L(\mathbf{y}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{2},$$

otherwise, classify into P_2.

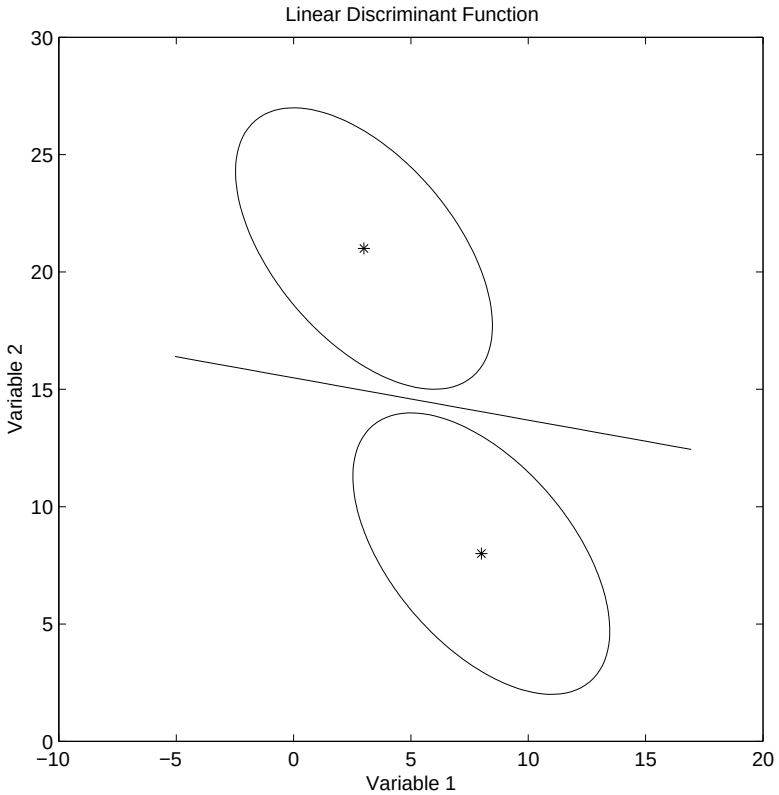
□

Note that the discriminant function is linear in $\mathbf{y}$. It is sometimes called Fisher's linear discriminant function. The function also is called $D(\mathbf{y})$ where the D stands for discriminant.

As an example, suppose that there are only $d = 2$ response measures. The response vector is distributed $\mathrm{N}(\boldsymbol{\mu}_i\boldsymbol{\Sigma})$ in population i, where

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 8 \\ 8 \end{pmatrix}; \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ 21 \end{pmatrix}; \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 5 & -3 \\ -3 & 6 \end{pmatrix}.$$

For simplicity, assume that $C(2|1) = C(1|2)$ and that $\pi_1 = \pi_2 = \frac{1}{2}$. Thus, $c = 0$. A display of the linear discriminant function appears below. The constant density ellipses include 95% of the underlying populations.



If the population covariances are not identical, then the discriminant function is quadratic in $\mathbf{y}$.

Theorem 10.3 (Two Population Quadratic Discrimination) *Suppose that*

$$\mathbf{y} \sim \begin{cases} \mathrm{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) & \text{if } P_1; \text{ and} \\ \mathrm{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) & \text{if } P_2. \end{cases}$$

Then, the Bayes rule is to classify into P_1 if

$$Q(\mathbf{y}) \geq c,$$

where

$$c = \ln\left(\frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}} \times \pi_2 \times C(1|2)}{|\boldsymbol{\Sigma}_2|^{\frac{1}{2}} \times \pi_1 \times C(2|1)}\right),$$

and

$$Q(\mathbf{y}) = (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{y} - \frac{\mathbf{y}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{y}}{2} - \frac{\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2}{2},$$

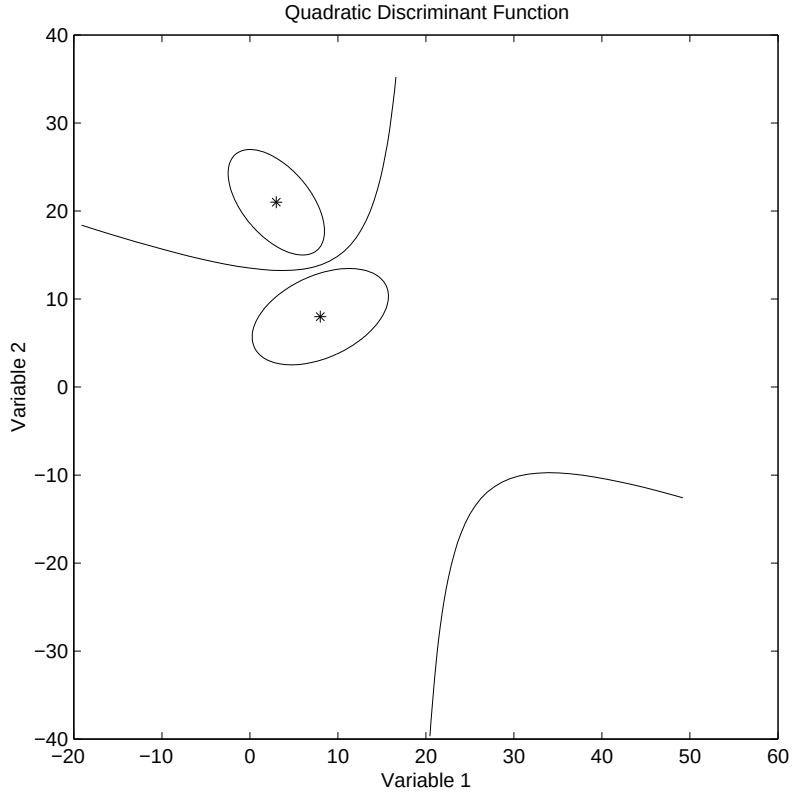
otherwise, classify into P_2.

□

As an example, suppose that there are only $d = 2$ response measures. The response vector is distributed $\mathrm{N}(\boldsymbol{\mu}_i\boldsymbol{\Sigma}_i)$ in population i, where

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 8 \\ 8 \end{pmatrix}; \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ 21 \end{pmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 10 & 3 \\ 3 & 5 \end{pmatrix}; \text{ and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & -3 \\ -3 & 6 \end{pmatrix}.$$

For simplicity, assume that $C(2|1) = C(1|2)$ and that $\pi_1 = \pi_2 = \frac{1}{2}$. Thus, $c = 0$. A display of the quadratic discriminant function appears below. The constant density ellipses include 95% of the underlying populations.



10.2.1 Probability of Misclassification

Suppose that

$$\mathbf{y} \sim \begin{cases} \mathrm{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) & \text{if } P_1; \text{ and} \\ \mathrm{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) & \text{if } P_2. \end{cases}$$

From the above results,

$$L(\mathbf{y}) \sim \begin{cases} \mathrm{N}\left(\frac{\delta}{2}, \delta\right) & \text{if } \mathbf{y} \text{ is from } P_1, \text{ and} \\ \mathrm{N}\left(-\frac{\delta}{2}, \delta\right) & \text{if } \mathbf{y} \text{ is from } P_2, \text{ where} \end{cases}$$
$$\delta = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

The parameter δ is called the squared Mahalanobis distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

The misclassification probabilities are

$$\Pr(2|1) = \Phi\left[\frac{c - \frac{\delta}{2}}{\sqrt{\delta}}\right],$$

and

$$\Pr(1|2) = 1 - \Phi\left[\frac{c + \frac{\delta}{2}}{\sqrt{\delta}}\right],$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

As an example, suppose that there are only $d = 2$ response measures. The response vector is distributed $\mathrm{N}(\boldsymbol{\mu}_i \boldsymbol{\Sigma})$ in population i, where

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 8 \\ 8 \end{pmatrix}; \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ 21 \end{pmatrix}; \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 5 & -3 \\ -3 & 6 \end{pmatrix}.$$

For simplicity, assume that $C(2|1) = C(1|2)$ and $\pi_1 = \pi_2 = \frac{1}{2}$. Thus, $c = 0$. In this case, $\delta = 28.8095$, $\delta/2 = 14.4048$, and $\Pr(2|1) = \Pr(1|2) = 0.0036$.

10.2.2 Minimax Rules

If the π_i are not known, then we cannot obtain the Bayes rule. However, we may still be able to find a minimax rule. A minimax rule is one which minimizes the maximum conditional risk. Finding minimax rules is sometimes simplified by using the following result:

Theorem 10.4 (Minimax) *If $(\mathcal{Y}_1, \mathcal{Y}_2)$ is a Bayes rule for some prior, and $(\mathcal{Y}_1, \mathcal{Y}_2)$ has constant risk, then $(\mathcal{Y}_1, \mathcal{Y}_2)$ is minimax.*

Proof: Suppose $(\mathcal{Y}_1, \mathcal{Y}_2)$ is the Bayes rule corresponding to the prior (π_1, π_2) and the risk is constant: $r(\mathcal{Y}_1, \mathcal{Y}_2|1) = r(\mathcal{Y}_1, \mathcal{Y}_2|2)$. Let $(\mathcal{Y}_1^*, \mathcal{Y}_2^*)$ be any other rule. Then

$$\begin{aligned} r(\mathcal{Y}_1, \mathcal{Y}_2) &= \pi_1 \times r(\mathcal{Y}_1, \mathcal{Y}_2|1) + \pi_2 \times r(\mathcal{Y}_1, \mathcal{Y}_2|2) \\ &= r(\mathcal{Y}_1, \mathcal{Y}_2|1) = r(\mathcal{Y}_1, \mathcal{Y}_2|2) = \max_i \; r(\mathcal{Y}_1, \mathcal{Y}_2|i), \end{aligned}$$

because risk is constant. Also, $(\mathcal{Y}_1, \mathcal{Y}_2)$ is Bayes, so that $r(\mathcal{Y}_1, \mathcal{Y}_2) \le r(\mathcal{Y}_1^*, \mathcal{Y}_2^*)$ must be satisfied. Note that

$$r(\mathcal{Y}_1, \mathcal{Y}_2) \le r(\mathcal{Y}_1^*, \mathcal{Y}_2^*) = \pi_1 \times r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|1) + \pi_2 \times r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|2) \le \max_i \; r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|i).$$

Thus,

$$\max_i \; r(\mathcal{Y}_1, \mathcal{Y}_2|i) \le \max_i \; r(\mathcal{Y}_1^*, \mathcal{Y}_2^*|i),$$

and $(\mathcal{Y}_1, \mathcal{Y}_2)$ is minimax.

□

To find the minimax rule, find the prior such that the Bayes rule

$$\mathcal{Y}_1 = \{\mathbf{y}; f_1(\mathbf{y}) \times C(2|1) \times \pi_1 \ge f_2(\mathbf{y}) \times C(1|2) \times \pi_2\}$$

results in

$$C(2|1) \times \Pr(2|1) = C(1|2) \times \Pr(1|2).$$

In the normal theory, $k = 2$, case with equal variances, the Bayes rule is to classify into P_1 if $L(\mathbf{y}) \ge c$. To obtain constant risk, choose c to satisfy

$$C(2|1) \times \Phi\left[\frac{c - \frac{\delta}{2}}{\sqrt{\delta}}\right] = C(1|2) \times \left(1 - \Phi\left[\frac{c + \frac{\delta}{2}}{\sqrt{\delta}}\right]\right).$$

The minimax rule is to classify $\mathbf{y}$ into P_1 if $L(\mathbf{y}) > c$ where c satisfies the above. In nonnormal problems, finding the minimax rule can be more difficult.

10.3 k-POPULATION CLASSIFICATION ANALYSIS

The optimal k-population classification rule can be obtained using the same approach as for 2 populations.

10.3.1 Optimal Classification Rule

Employing the same notation as for $k = 2$, the risk, using the classification rule $(\mathcal{Y}_1, \ldots, \mathcal{Y}_k)$, is

$$r(\mathcal{Y}_1, \ldots, \mathcal{Y}_k) = \sum_{i=1}^{k} \pi_i \sum_{j=1}^{k} \left[C(j|i) \Pr(j|i, \mathcal{Y}_1, \ldots, \mathcal{Y}_k) \right]$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{k} \left[\pi_i C(j|i) \Pr(j|i, \mathcal{Y}_1, \ldots, \mathcal{Y}_k) \right] = \sum_{j=1}^{k} \sum_{i=1}^{k} \left[\pi_i C(j|i) \int_{\mathcal{Y}_j} f_i(\mathbf{y})\, d\mathbf{y} \right]$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{k} \int_{\mathcal{Y}_j} \left[\pi_i C(j|i) f_i(\mathbf{y}) \right] d\mathbf{y} = \sum_{j=1}^{k} \int_{\mathcal{Y}_j} h_j(\mathbf{y})\, d\mathbf{y}, \text{ where}$$

$$h_j(\mathbf{y}) = \sum_{i=1}^{k} \left[\pi_i C(j|i) f_i(\mathbf{y}) \right].$$

The risk also can be written as

$$r(\mathcal{Y}_1, \ldots, \mathcal{Y}_k) = \int_{\mathcal{Y}} \sum_{j=1}^{k} \left[\delta_j(\mathbf{y}) h_j(\mathbf{y}) \right] d\mathbf{y} \text{ where } \delta_j(\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{y} \in \mathcal{Y}_j; \\ 0 & \text{otherwise.} \end{cases}$$

To minimize the risk, consider minimizing

$$Q(\mathbf{y}) = \sum_{j=1}^{k} \left[\delta_j(\mathbf{y}) h_j(\mathbf{y}) \right]$$

for each $\mathbf{y} \in \mathcal{Y}$, subject to the restriction that $\delta_j(\mathbf{y})$ must equal 1 for exactly one value of j. That is, an observation must be classified into one and only one population. Suppose that the values of $h_j(\mathbf{y})$ for $i = 1, \ldots, k$ are distinct. In this case, the quantity $Q(\mathbf{y})$ can be minimized by assigning the values

$$\delta_j(\mathbf{y}) = \begin{cases} 1 & \text{if } h_j(\mathbf{y}) = \min_i h_i(\mathbf{y}), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

If there are ties for $\min_i h_i(\mathbf{y})$, then the the observation can be assigned to any population for which $h_j(\mathbf{y})$ is a minimum. The result is summarized in the following theorem.

Theorem 10.5 (k Population Rule) *The k population Bayes classification rule is to assign $\mathbf{y}$ to P_m if $h_m(\mathbf{y}) \leq h_j(\mathbf{y})$ for all $j \neq m$, where*

$$h_j(\mathbf{y}) = \sum_{i=1}^{k} \left[\pi_i C(j|i) f_i(\mathbf{y}) \right].$$

□

The rule k population rule simplifies when all costs are equal:

$$C(i|j) = \begin{cases} k & \text{if } i \neq j; \\ 0 & \text{if } i = j. \end{cases}$$

Let

$$f(\mathbf{y}) = \sum_{i=1}^{k} [\pi_i f_i(\mathbf{y})].$$

Note that $f(\mathbf{y})$ is a density function. It integrates to 1 and is nonnegative. In particular, it is the density function of a random vector randomly selected from population i with probability π_i. This type of density is often called a mixture. Note that, when all costs are equal, $h_j(\mathbf{y}) = f(\mathbf{y}) - \pi_j f_j(\mathbf{y})$. Hence, the decision rule simplifies to the following. Classify $\mathbf{y}$ to P_m if $\pi_m f_m(\mathbf{y}) \geq \pi_j f_j(\mathbf{y})$ for all $j \neq m$. This rule is equivalent to assigning $\mathbf{y}$ to the population having the highest posterior probability.

Prior to observing $\mathbf{y}$, the probability that an observation belongs to the i^{th} population is π_i. It is instructive to examine how that probability is modified after observing the data. That is, if $\Pr(P_i) = \pi_i$, then what is $\Pr(P_i|\mathbf{y})$?

Theorem 10.6 (Posterior Probability) *The posterior probability that* $\mathbf{y}$ *was sampled from* P_i *is*

$$\Pr(P_i|\mathbf{y}) = \frac{\pi_i f_i(\mathbf{y})}{\sum_{j=1}^{k} \pi_j f_j(\mathbf{y})} = \frac{\pi_i f_i(\mathbf{y})}{f(\mathbf{y})}.$$

Proof:

$$\Pr(P_i|\mathbf{y}) = \frac{\Pr(P_i, \mathbf{y})}{f(\mathbf{y})} = \frac{\Pr(\mathbf{y}|P_i)\Pr(P_i)}{f(\mathbf{y})} = \frac{\pi_i f_i(\mathbf{y})}{f(\mathbf{y})}.$$

□

10.3.2 k Normal Populations

For k normally distributed populations having equal covariance matrices, the Bayes decision rule is to classify into P_m if $h_m(\mathbf{y}) \leq h_j(\mathbf{y})$ for all j, where

$$h_j(\mathbf{y}) = \sum_{i=1}^{k} \pi_i C(j|i) \frac{\exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}_i)\right\}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}}.$$

For simplicity, assume equal costs of misclassification:

$$C(i|j) = \begin{cases} k & \text{if } i \neq j; \\ 0 & \text{if } i = j. \end{cases}$$

Then, the Bayes rule is to classify $\mathbf{y}$ into P_i if $\pi_i f_i(\mathbf{y}) \geq \pi_j f_j(\mathbf{y})$ for all j. That is, classify $\mathbf{y}$ into P_i if

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)}{2} - \ln\left(\frac{\pi_j}{\pi_i}\right) \geq 0$$

for all j. As shown above, this rule is equivalent to classifying the observation into the population having the largest posterior probability. In the normal case, the posterior probability is

$$\Pr(P_i|\mathbf{y}) = \frac{e^{-\frac{1}{2}D_i^2(\mathbf{y})}}{\sum_{j=1}^{k} e^{-\frac{1}{2}D_j^2(\mathbf{y})}},$$

where

$$D_i^2(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}_i) - 2\ln(\pi_i).$$

Note, the population with the highest posterior probability is the one with the smallest value of $D_i^2(\mathbf{y})$. Excluding the $-2\ln(\pi_i)$ term, $D_i^2(\mathbf{y})$ is called the squared Mahalanobis distance between $\mathbf{y}$ and $\boldsymbol{\mu}_i$.

10.4 SELECTION OF VARIABLES (2 GROUPS)

Suppose a random sample of size n_1 is available from $\mathrm{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and a random sample of size n_2 is available from $\mathrm{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. The corresponding linear model is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \end{pmatrix},$$

$\mathbf{Y}$ is $N \times d$ and $N = n_1 + n_2$. We are interested in knowing whether a subset of the d variables (say the first r) can discriminate as well as the entire set. The coefficients of the linear discriminant function are

$$\boldsymbol{\tau} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Partition $\mathbf{Y}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ so that $\mathbf{Y}_1$ is $n \times r$, $\boldsymbol{\Sigma}_{11}$ is $r \times r$, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\tau}_1$ are each $r \times 1$. Then, the last $d - r$ variables add nothing to the discriminant function if $\boldsymbol{\tau}_2 = 0$. Thus, we wish to test $\mathrm{H}_0 \colon \boldsymbol{\tau}_2 = \mathbf{0}$. Use partitioned matrix results to show that

$$\boldsymbol{\tau} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} \Longrightarrow \boldsymbol{\tau}_2 = \boldsymbol{\Sigma}_{22\cdot 1}^{-1}(\boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1).$$

Note that

$$\boldsymbol{\tau}_2 = \mathbf{0} \Longleftrightarrow \boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1 = \mathbf{0}.$$

Thus, we are interested in testing $\mathrm{H}_0 \colon \boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1 = \mathbf{0}$.

Theorem 10.7 *Let*

$$T^2 = \left(\frac{n_1 n_2}{N}\right) \hat{\boldsymbol{\theta}}' \mathbf{S}^{-1} \hat{\boldsymbol{\theta}},$$

where $\mathbf{S} = \mathbf{Y}'(\mathbf{I} - \mathbf{M})\mathbf{Y}/(N-2)$, $\mathbf{M} = \mathrm{ppo}(\mathbf{X})$, *and* $\hat{\boldsymbol{\theta}} = \overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2$. *Also, let*

$$T_1^2 = \left(\frac{n_1 n_2}{N}\right) \hat{\boldsymbol{\theta}}_1' \mathbf{S}_{11}^{-1} \hat{\boldsymbol{\theta}}_1,$$

where $\hat{\boldsymbol{\theta}}_1$ *is* $r \times 1$ *(i.e., the first partition of* $\hat{\boldsymbol{\theta}}$*) and* $\mathbf{S}_{11}$ *is the upper left hand* $r \times r$ *block of* $\mathbf{S}$. *Then an* α *size test of* H_0 *is to reject* H_0 *if* $F_2 \geq F_{d-r,N-d-1}^{1-\alpha}$, *where*

$$F_2 = \frac{T^2 - T_1^2}{1 + (N-2)^{-1}T_1^2}\left(\frac{N-d-1}{(d-r)(N-2)}\right).$$

Conditional on $\mathbf{Y}_1$, F_2 *is distributed as*

$$F_2 | \mathbf{Y}_1 \sim F_{d-r,N-d-1,\lambda},$$

where

$$\lambda = \left(\frac{n_1 n_2}{N}\right) \frac{(\boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1)' \boldsymbol{\Sigma}_{22\cdot 1}^{-1} (\boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1)}{2[1 + (N-2)^{-1}T_1^2]}.$$

Proof: See the description of Roy's step-down tests in these notes.

□

Remark — If $\boldsymbol{\tau}_2 = \mathbf{0}$, then F_2 has an unconditional central F distribution. The above testing procedure is the basis of stepwise tests in discriminant function analysis. The procedure can be extended to more than two groups (see Seber, page 341). However, the test is not recommended.

10.4.1 Alternative Approach to Variable Selection

As in the previous section, a test of $\boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1 = \mathbf{0}$ is desired. Write the two sample model as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \end{pmatrix}, \text{ and } \operatorname{vec}(\mathbf{U}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n).$$

Consider the distribution of $\mathbf{Y}_2$, conditional on $\mathbf{Y}_1$:

$$\operatorname{vec}(\mathbf{Y}_2)|\mathbf{Y}_1 \sim \mathrm{N}_{d-r}\left[\operatorname{vec}\left(\mathbf{XB}_{2\cdot1} + \mathbf{Y}_1\boldsymbol{\Gamma}\right), \left(\boldsymbol{\Sigma}_{22\cdot1} \otimes \mathbf{I}_n\right)\right],$$

where

$$\mathbf{B}_{2\cdot1} = \begin{pmatrix} \boldsymbol{\mu}_{12}' \\ \boldsymbol{\mu}_{22}' \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu}_{11}' \\ \boldsymbol{\mu}_{21}' \end{pmatrix} \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, \quad \boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12},$$

and where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ have been partitioned conformably to $\boldsymbol{\theta}$ as

$$\boldsymbol{\mu}_1 = \begin{pmatrix} \boldsymbol{\mu}_{11} \\ \boldsymbol{\mu}_{12} \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{pmatrix} \boldsymbol{\mu}_{21} \\ \boldsymbol{\mu}_{22} \end{pmatrix}.$$

Let

$$\mathbf{W} = \begin{pmatrix} \mathbf{X} & \mathbf{Y}_1 \end{pmatrix} \text{ and let } \mathbf{G} = \begin{pmatrix} \mathbf{B}_{2\cdot1} \\ \boldsymbol{\Gamma} \end{pmatrix}.$$

Then, the conditional model can be written as

$$\mathbf{Y}_2 = \mathbf{WG} + \mathbf{U}_{2\cdot1},$$

where

$$\operatorname{vec}(\mathbf{U}_{2\cdot1}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{22\cdot1} \otimes \mathbf{I}_n).$$

Let $\mathbf{c}$ be an $(r+2) \times 1$ vector defined as

$$\mathbf{c} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Note that

$$\mathbf{c}'\mathbf{G} = \begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{B}_{2\cdot1} = \left[\boldsymbol{\theta}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\theta}_1\right]'.$$

The likelihood ratio test of $\mathrm{H}_0\colon \mathbf{c}'\mathbf{G} = \mathbf{0}$ in the conditional model is to reject H_0 for small

$$U_{2\cdot1} = \frac{|\mathbf{E}_{22\cdot1}|}{|\mathbf{E}_{22\cdot1} + \mathbf{H}_{22\cdot1}|},$$

where

$$\mathbf{E}_{22\cdot1} = \mathbf{Y}_2'(\mathbf{I}_n - \mathbf{H}_w)\mathbf{Y}_2, \quad \mathbf{H}_w = \operatorname{ppo}(\mathbf{W}),$$

$$\mathbf{H}_{22\cdot1} = \widehat{\mathbf{G}}'\mathbf{c}\left[\mathbf{c}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{c}\right]^{-1}\mathbf{c}'\widehat{\mathbf{G}}, \text{ and } \widehat{\mathbf{G}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}_2.$$

It can be shown that this test is identical to the test based on differences between T^2 statistics. Extra credit for proof of the equivalence.

10.5 KERNEL-BASED CLASSIFICATION

Generally the densities $f_i(\mathbf{y})$, for $i = 1, \ldots, k$ are unknown and must be estimated. If the parametric family is known, then maximum likelihood can be used to estimate the unknown parameters. If the parametric family is unknown, then other procedures can be used. One such procedure yields kernel density estimators.

Let G_i be the training sample from population i. That is,

$$G_i = \{\mathbf{y}_j; \mathbf{y}_j \in \text{ sample from population } i\}.$$

A kernel density estimator for $f_i(\mathbf{z})$ can be written as

$$\widehat{f}_i(\mathbf{z}) = \frac{1}{n_i} \sum_{\mathbf{y} \in G_i} K_i(\mathbf{z} - \mathbf{y}, r),$$

where $K_i(\mathbf{u})$ is a pdf and r is a smoothing parameter. For example, the kernel for a p-variate normal density is

$$K_i(\mathbf{u}, r_i) = \frac{\exp\left\{-\frac{1}{2r_i^2}\mathbf{u}'\boldsymbol{\Sigma}_i^{-1}\mathbf{u}\right\}}{r_i^p (2\pi)^{p/2} |\boldsymbol{\Sigma}_i|^{1/2}}$$

and the kernel for the p-dimensional uniform density is

$$K_i(\mathbf{u}, r_i) = \begin{cases} \frac{\Gamma\left(\frac{p}{2}+1\right)}{r_i^p |\boldsymbol{\Sigma}_i|^{\frac{1}{2}} \pi^{\frac{p}{2}}} & \text{if } \mathbf{u}'\boldsymbol{\Sigma}_i^{-1}\mathbf{u} \le r_i^2 \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

The quantity $\pi^{p/2}\Gamma(p/2+1)$ is the volume of a p-dimensional unit radius sphere. The quantity $\Gamma(p/2+1)/(r_i^p|\boldsymbol{\Sigma}_i|^{1/2}\pi^{p/2})$ is the volume of a p-dimensional ellipsoid bounded by $\{\mathbf{u}|\mathbf{u}'\boldsymbol{\Sigma}_i^{-1}\mathbf{u}\} = r_i^2$ Many other kernels could be used as well.

The matrices $\boldsymbol{\Sigma}_i$ for $i = 1, \ldots, k$ can be estimated in many ways, depending on which assumptions are reasonable. For example, if it can be assumed that population covariance matrices are homogeneous, then $\boldsymbol{\Sigma}_i$ can be estimated by $\mathbf{Y}'(\mathbf{I} - \mathbf{H}_x)\mathbf{Y}/(n-k)$ for each i, where $\mathbf{X}$ is the design matrix for a one-way classification with k treatments and $\mathbf{Y}$ is the total training sample. If covariance matrices are heterogeneous, then $\boldsymbol{\Sigma}_i$ can be estimated by $\mathbf{Y}_i'(\mathbf{I} - \mathbf{H}_i)\mathbf{Y}_i/(n_i - 1)$, where $\mathbf{H}_i = \text{ppo}(\mathbf{1}_{n_i})$ and $\mathbf{Y}_i$ is the training sample from population i.

Reasonable values for r_i are

$$r_i = \left(\frac{4}{(2p+1)n_i}\right)^{\frac{1}{p+4}} \text{ if a normal kernel is employed, and}$$

$$r_i = \left(\frac{2^{p+2}(p+2)\Gamma\left(\frac{p}{2}\right)}{n_i p}\right)^{\frac{1}{p+4}} \text{ if a uniform kernel is employed.}$$

10.6 NEAREST NEIGHBOR CLASSIFICATION

Nearest neighbor classification consists of using the Bayes rule but substituting nearest neighbor density estimators for the unknown densities. To construct a nearest neighbor density estimator for $f_i(\mathbf{z})$, first choose an integer number of neighbors to be examined, g. Several values of g can be tried and values near $\sqrt{n/k}$ are sensible starting points. For a fixed $\mathbf{z}$, find the g nearest neighbors in the training set. The squared distance from $\mathbf{z}$ to $\mathbf{y}$ is measured as

$$D^2(\mathbf{y}, \mathbf{z}) = (\mathbf{z} - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{z} - \mathbf{y}),$$

where $\mathbf{S}$ is the pooled sample covariance matrix. Let $\mathbf{y}_g$ be the g^{th} nearest neighbor to $\mathbf{z}$ and define $r(\mathbf{z})$ as

$$r(\mathbf{z}) = \left[(\mathbf{z} - \mathbf{y}_g)'\mathbf{S}^{-1}(\mathbf{z} - \mathbf{y}_g)\right]^{\frac{1}{2}}.$$

The volume of the p-dimensional ellipsoid

$$A(\mathbf{z}) = \{\boldsymbol{\ell}|(\boldsymbol{\ell} - \mathbf{z})'\mathbf{S}^{-1}(\boldsymbol{\ell} - \mathbf{z}) \le r(\mathbf{z})^2\}$$

is

$$V(\mathbf{z}) = \frac{r(\mathbf{z})^p \pi^{\frac{p}{2}} |\mathbf{S}|^{\frac{1}{2}}}{\Gamma\left(\frac{p}{2}+1\right)}.$$

The nearest neighbor density estimator is

$$\widehat{f}_i(\mathbf{z}) = \frac{g_i}{n_i V(\mathbf{z})},$$

where g_i is the number of the g nearest neighbors that belong to population i.

Note that if $\mathbf{Z}$ is a random p vector from population i, $\mathbf{z}$ is a fixed p-vector, and $A(\mathbf{z})$ is an ellipsoid centered at $\mathbf{z}$, then

$$\frac{P\left[\mathbf{Z} \in A(\mathbf{z})|i\right]}{V(\mathbf{z})} = \frac{\int_{A(\mathbf{z})} f_i(\mathbf{u})\, d\mathbf{u}}{V(\mathbf{z})} \text{ and}$$

$$\lim_{V(\mathbf{z})\to 0} \frac{P\left[\mathbf{Z} \in A(\mathbf{z})|i\right]}{V(\mathbf{z})} = \lim_{V(\mathbf{z})\to 0} \frac{V(\mathbf{z}) f_i(\mathbf{z})}{V(\mathbf{z})} = f_i(\mathbf{z}),$$

provided that the density is sufficiently smooth. Accordingly, the kernel density estimator is consistent.

If misclassification costs are equal, then the Bayes rule is to classify $\mathbf{z}$ to the population with the largest posterior probability. The posterior probability is

$$P(P_i|\mathbf{z}) = \frac{f_i(\mathbf{z})\pi_i}{f(\mathbf{z})}$$

and is estimated by

$$\widehat{P}(P_i|\mathbf{z}) = \frac{\pi_i \dfrac{g_i}{n_i V(\mathbf{z})}}{\displaystyle\sum_{j=1}^{k} \pi_j \frac{g_j}{n_j V(\mathbf{z})}} = \frac{\pi_i \dfrac{g_i}{n_i}}{\displaystyle\sum_{j=1}^{k} \pi_j \frac{g_j}{n_j}}.$$

Furthermore, if priors are proportional to sample size ($\pi_i = n_i/n$), then

$$\widehat{P}(P_i|\mathbf{z}) = \frac{g_i}{g}.$$

10.7 LOGISTIC DISCRIMINATION

In this section, we assume equal costs of misclassification:

$$C(i|j) = \begin{cases} 1 & \text{if } i \neq j; \\ 0 & \text{if } i = j. \end{cases}$$

Accordingly, the Bayes rule is to classify $\mathbf{y}$ into the population having the largest posterior probability:

$$\Pr(P_i|\mathbf{y}) = \frac{\pi_i f_i(\mathbf{y})}{\displaystyle\sum_{j=1}^{k} \pi_j f_j(\mathbf{y})}.$$

Suppose, as is usual, that $f_i(\mathbf{y})$ must be estimated from a training sample. In the k population Gaussian case with equal variances, $d(d+2k+1)/2$ parameters must be estimated.

The logistic approach focuses on estimating the posterior probability directly, rather than estimating the densities. A particular form is not assumed for the density functions. Rather, the linear logistic approach assumes that

$$\ln\left(\frac{f_i(\mathbf{y})}{f_k(\mathbf{y})}\right) = \alpha_i + \boldsymbol{\beta}_i'\mathbf{y},$$

for $i = 1, \ldots, k-1$. Accordingly, the posterior probabilities can be written as

$$\Pr(P_i|\mathbf{y}) = \frac{\left(\frac{\pi_i f_i(\mathbf{y})}{\pi_k f_k(\mathbf{y})}\right)}{\sum_{j=1}^{k}\left(\frac{\pi_j f_j(\mathbf{y})}{\pi_k f_k(\mathbf{y})}\right)} = \frac{\left(\frac{\pi_i f_i(\mathbf{y})}{\pi_k f_k(\mathbf{y})}\right)}{1+\sum_{j=1}^{k-1}\left(\frac{\pi_j f_j(\mathbf{y})}{\pi_k f_k(\mathbf{y})}\right)}$$

$$= \begin{cases} \frac{\exp\left[\ln\left(\frac{\pi_i}{\pi_k}\right)+\alpha_i+\boldsymbol{\beta}_i'\mathbf{y}\right]}{1+\sum_{j=1}^{k-1}\exp\left[\ln\left(\frac{\pi_j}{\pi_k}\right)+\alpha_j+\boldsymbol{\beta}_j'\mathbf{y}\right]} & \text{if } i \neq k; \text{ and} \\ \frac{1}{1+\sum_{j=1}^{k-1}\exp\left[\ln\left(\frac{\pi_j}{\pi_k}\right)+\alpha_j+\boldsymbol{\beta}_j'\mathbf{y}\right]} & \text{if } i = k. \end{cases}$$

The approach is called linear logistic because, for $k = 2$, the logit (log odds) is linear in $\mathbf{y}$:

$$\ln\left(\frac{\Pr(P_1|\mathbf{y})}{1-\Pr(P_1|\mathbf{y})}\right) = \ln\left(\frac{\pi_1}{\pi_2}\right) + \alpha + \boldsymbol{\beta}'\mathbf{y}.$$

In the k population case, $(d+1)(k-1)$ parameters must be estimated. Denote the vector containing the entire set of parameters by $\boldsymbol{\Theta}$. Assuming that an independent sample of size n_i has been obtained from the i^{th} population for $i = 1, \ldots, k$. The likelihood function for the entire sample is

$$L(\boldsymbol{\Theta}|\mathbf{Y}) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\Pr(P_i|\mathbf{y}_{ij}),$$

where the posterior probabilities are given above. The log likelihood function can be maximized by Newton methods.

Parameter estimation in linear logistic discrimination analysis with $k = 2$ is identical to parameter estimation in logistic regression. This is a Stat 539 topic.

Chapter 11

PRINCIPAL COMPONENTS

11.1 POPULATION PRINCIPAL COMPONENTS

Principal components can be motivated in a variety of ways. We will examine two motivations. The most straightforward motivation considers the variance maximizing properties of principal components. The second motivation considers the dimension reduction properties of principal components.

11.1.1 Maximizing the Variance of Linear Combinations

Let $\mathbf{y}$ be a random d-vector with distribution $\mathbf{y} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose that the experimenter does not wish to use the original d variables because it is believed that a smaller set of linear combinations of the d variables will capture most of the information. A linear combination is said to have no information if all members of the population have exactly the same score on the combination. That is, the linear combination is not informative if, with probability 1, $\mathbf{h}'\mathbf{y} = c$. In this case, $\text{var}(\mathbf{h}'\mathbf{y}) = 0$. On the other hand, a linear combination is said to have much information if the members of the population vary greatly on their scores. In this case, $\text{var}(\mathbf{h}'\mathbf{y})$ is large. A sensible goal is to keep linear combinations which have large variance and discard linear combinations which have small variance.

The first (most important) linear combination is called the first principal component and is $z_1 = \mathbf{h}'(\mathbf{y} - \boldsymbol{\mu})$ where $\mathbf{h}$ is chosen such that

$$Q(\mathbf{h}) = \text{var}\left(\frac{\mathbf{h}'(\mathbf{y} - \boldsymbol{\mu})}{\sqrt{\mathbf{h}'\mathbf{h}}}\right)$$

is maximized. The second principal component is $z_2 = \mathbf{h}'(\mathbf{y} - \boldsymbol{\mu})$ where $\mathbf{h}$ is chosen to maximize $Q(\mathbf{h})$, subject to $\text{cov}(z_2, z_1) = 0$. In general, the k^{th} principal component is $z_k = \mathbf{h}'(\mathbf{y} - \boldsymbol{\mu})$ where $\mathbf{h}$ is chosen to maximize $Q(\mathbf{h})$, subject to $\text{cov}(z_k, z_i) = 0$ for $i = 1, \ldots, k-1$.

Theorem 11.1 *Write* $\boldsymbol{\Sigma}$ *in diagonal form:* $\boldsymbol{\Sigma} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$ *where* $\mathbf{T} = \begin{pmatrix}\mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_d\end{pmatrix}$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$, *and* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. *The* i^{th} *principal component is* $z_i = \mathbf{t}_i'(\mathbf{y} - \boldsymbol{\mu})$ *and* $\text{var}(z_i) = \lambda_i$, *for* $i = 1, \ldots, d$.

11.1.2 Dimension Reduction Properties (Optional)

This section is from Seber (1984, p. 176–181). Consider the following problem. We wish to find a random vector $\mathbf{z}$: $k \times 1$ and a matrix of constants $\mathbf{A}$: $d \times k$ for $k < d$ such that the vector $\mathbf{Az}$ is "close" to $\mathbf{y} - \boldsymbol{\mu}$. To measure closeness, let $\mathbf{u} = (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{Az}$. Denote $\text{var}(\mathbf{u})$ by $\boldsymbol{\Sigma}_u$. Then, we will say that $\mathbf{Az}$ is close to $(\mathbf{y} - \boldsymbol{\mu})$ if $\boldsymbol{\Sigma}_u$ is small. The magnitude of $\boldsymbol{\Sigma}_u$ will be indexed by $f(\boldsymbol{\Sigma}_u)$ for some function f defined on the space of all $d \times d$ positive semidefinite matrices.

What f should be used? Let $\mathbf{S}$ and $\mathbf{V}$ be psd matrices of the same order. We wish to determine the relative magnitudes of $\mathbf{S}$ and $\mathbf{V}$. A "reasonable" function, f, ought to satisfy the following two properties:

(i) If $\mathbf{S} \neq \mathbf{V}$, and $\mathbf{S} - \mathbf{V} \geq 0$, then $f(\mathbf{S}) > f(\mathbf{V})$.

(ii) If $\mathbf{Q}$ is an orthogonal matrix, then $f(\mathbf{S}) = f(\mathbf{QSQ}')$.

Point (ii) above stems from the following considerations. Premultiplication of $\mathbf{u}$ by an orthogonal matrix $\mathbf{Q}$ performs a perpendicular rotation of the d axes. The relative magnitudes of $\mathbf{u}$ have not been changed so it is sensible to expect that $f(\mathbf{\Sigma}_u) = f(\mathbf{Q}\mathbf{\Sigma}_u\mathbf{Q}')$.

The above two conditions ought to help us narrow down the sort of function, f, to employ. In fact, the two conditions are quite helpful in restricting f. Before proceeding, we will pick up a couple of useful results.

Theorem 11.2 (Courant Fischer Min-Max Theorem) *Let* $\mathbf{L}\colon d \times k$ *be a rank-k matrix, and let* $\mathbf{W}$ *be a* $d \times d$ *psd matrix with characteristic roots* $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_d$*. Then*

$$\inf_{\mathbf{L}} \sup_{\mathbf{c};\mathbf{c}'\mathbf{L}=\mathbf{0}} \frac{\mathbf{c}'\mathbf{W}\mathbf{c}}{\mathbf{c}'\mathbf{c}} = \theta_{k+1}.$$

Proof: See page 525 in Seber.

□

Corollary to Theorem 11.2 Let $\mathbf{S}$ and $\mathbf{V}$ be $d \times d$ psd matrices. Denote the ordered ch. roots of $\mathbf{S}$ by $r_i(\mathbf{S})$ for $i = 1, \ldots, d$; and denote the ordered ch. roots of $\mathbf{V}$ by $r_i(\mathbf{V})$ for $i = 1, \ldots, d$. Then, $\mathbf{S} - \mathbf{V} \geq 0 \Rightarrow r_i(\mathbf{S}) \geq r_i(\mathbf{V})$ for $i = 1, \ldots, d$.

Theorem 11.3 *The necessary and sufficient conditions for* $f(\cdot)$ *to satisfy (i) and (ii) are that* $f(\mathbf{A})$ *is a function of the roots of* $\mathbf{A}$ *and is strictly increasing in each argument.*

Proof: See page 177 in Seber. We will discuss the proof if time permits. The proof uses the Corollary to the Courant Fischer Theorem.

□

Theorem 11.4 *The matrix* $\mathbf{A}$ *and the random vector* $\mathbf{z}$ *which minimize* $f(\mathbf{\Sigma}_u)$ *are given by* $\mathbf{z} = \mathbf{T}_1'\mathbf{y}$ *and* $\mathbf{A} = \mathbf{T}_1$ *where* $\mathbf{\Sigma} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}'$, $\mathbf{T} = \begin{pmatrix}\mathbf{T}_1 & \mathbf{T}_2\end{pmatrix}$, $\mathbf{T}_1$ *is* $d \times k$*, and the columns of* $\mathbf{T}_1$ *are the characteristic vectors corresponding to the k largest roots of* $\mathbf{\Sigma}$*.*

Proof: In class if there is time.

□

11.2 INFERENCE ON PRINCIPAL COMPONENTS UNDER NORMALITY

Let $\mathbf{Y}$ be an $N \times d$ random matrix with distribution

$$\mathrm{vec}(\mathbf{Y}) \sim \mathrm{N}\left[\mathrm{vec}(\mathbf{X}\mathbf{B}), \mathbf{\Sigma} \otimes \mathbf{I}_n\right],$$

where $\mathbf{X}$ is a known $N \times p$ matrix of constants having rank r. Then

$$n\mathbf{S} \sim \mathrm{W}_d(n, \mathbf{\Sigma}), \text{ where } \mathbf{S} = n^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_x)\mathbf{Y}, \quad n = N - r,;$$

and $\mathbf{H}_x = \mathrm{ppo}(\mathbf{X})$. Denote the sorted eigenvalues of $\mathbf{\Sigma}$ by $\lambda_1 > \lambda_2 > \cdots > \lambda_d$ (assume that all eigenvalues are distinct). Denote the sorted eigenvalues of $\mathbf{S}$ by $\ell_1 > \ell_2 > \cdots > \ell_d$. Denote the corresponding normalized eigenvectors by $\boldsymbol{\gamma}_j$ and $\mathbf{g}_j$ for $j = 1, \ldots, d$. That is,

$$\mathbf{S}\mathbf{g}_j = \mathbf{g}_j\ell_j \text{ and } \mathbf{\Sigma}\boldsymbol{\gamma}_j = \boldsymbol{\gamma}_j\lambda_j.$$

Below are some distributional results for the sample eigenvalues and eigenvectors. These results depend heavily on the multivariate normality assumption.

Theorem 11.5 (LR test that the Variables are Uncorrelated) *Sometimes a test of* $\mathrm{H}_0\colon \mathbf{R} = \mathbf{I}$ *against* $\mathrm{H}_a\colon \mathbf{R} \neq \mathbf{I}$ *is desired, where* $\mathbf{R}$ *is the population correlation matrix. Under* H_0*, the statistic*

$$w = -\left[n - \frac{2d+5}{6}\right]\left[\ln|\mathbf{S}| - \sum_{j=1}^{d}\ln(s_{jj})\right]$$

is asymptotically distributed as a χ^2 *random variable with* $d(d-1)/2$ *degrees of freedom, where* $n = N - r$ *and* s_{jj} *is the* j^{th} *diagonal entry of* $\mathbf{S}$*.* □

Theorem 11.6 (Asymptotic Distribution of Eigenvalues and Eigenvectors) *Asymptotically, $\ell_1, \ldots, \ell_d$ and $\mathbf{G}$ are independently distributed as follows:*

$$\sqrt{n}(\ell_i - \lambda_i) \xrightarrow{\text{dist}} \text{N}\left(0, 2\lambda_i^2\right) \text{ for } i = 1, \ldots, d;$$

and $\sqrt{n}\,\text{vec}(\mathbf{G} - \boldsymbol{\Gamma})$ is asymptotically normal, where $n = N - r$. The specific distribution for $\mathbf{G}$ is

$$\sqrt{n}(\mathbf{g}_i - \boldsymbol{\gamma}_i) \xrightarrow{\text{dist}} \text{N}\left[\mathbf{0}, \sum_{j \neq i} \frac{\lambda_i \lambda_j \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j'}{(\lambda_i - \lambda_j)^2}\right] \text{ and } n\,\text{Cov}(\mathbf{g}_i, \mathbf{g}_j) \xrightarrow{\text{Prob}} -\frac{\lambda_i \lambda_j \boldsymbol{\gamma}_i \boldsymbol{\gamma}_j'}{(\lambda_i - \lambda_j)^2}.$$

Also, the natural log of ℓ_j is approximately normal:

$$\ln(\ell_j) \dot{\sim} \text{N}\left[\ln(\lambda_j), \frac{2}{n}\right].$$

A large sample 95% confidence interval for λ_j is given by

$$\left(\ell_j e^{-1.96\sqrt{2/n}}, \ell_j e^{1.96\sqrt{2/n}}\right).$$

Proof: See Anderson (1984) or Flury (Common Principal Components, 1988, John Wiley).

□

The results in Theorem 11.6 can be written several ways. The following corollary gives a matrix expression which is useful when making inferences on the vector $\boldsymbol{\lambda}$ or on the matrix $\boldsymbol{\Gamma}$.

Corollary to Theorem 11.6 The results in Theorem 11.6 can be summarized as follows: $\sqrt{n}(\boldsymbol{\ell} - \boldsymbol{\lambda})$ and $\sqrt{n}\,\text{vec}(\mathbf{G} - \boldsymbol{\Gamma})$ are asymptotically independent with distributions

$$\sqrt{n}(\boldsymbol{\ell} - \boldsymbol{\lambda}) \xrightarrow{\text{dist}} \text{N}\left(\mathbf{0}, 2\boldsymbol{\Lambda}^2\right) \text{ and } \sqrt{n}\,\text{vec}(\mathbf{G} - \boldsymbol{\Gamma}) \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \boldsymbol{\Theta}), \text{ where}$$

$$\boldsymbol{\Theta} = (\mathbf{I}_d \otimes \boldsymbol{\Gamma})\mathbf{V}(\mathbf{I}_d \otimes \boldsymbol{\Gamma})'; \quad \mathbf{V} = \{\mathbf{V}_{ij}\}; \quad \mathbf{V}_{ij} \text{ has dimension } d \times d;$$

$$\mathbf{V}_{ij} = \begin{cases} \text{diag}(\tau_{i1}, \ldots, \tau_{id}) & \text{if } i = j; \\ -\tau_{ij}\mathbf{e}_j\mathbf{e}_i' & \text{otherwise;} \end{cases}; \quad \tau_{ij} = \begin{cases} 0 & \text{if } i = j; \\ \dfrac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} & \text{otherwise;} \end{cases};$$

and $\mathbf{e}_i$ is the i^{th} column of $\mathbf{I}_d$.

Theorem 11.7 (Distribution of Variance Accounted for by the Smallest Eigenvalues) *Denote the proportion of the variance that the smallest q components account for by δ_q and denote the corresponding sample quantity by $\widehat{\delta}_q$. That is*

$$\delta_q = \frac{\sum_{i=d-q+1}^{d} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \text{ and } \widehat{\delta}_q = \frac{\sum_{i=d-q+1}^{d} \ell_i}{\sum_{i=1}^{d} \ell_i}.$$

If δ_q is small, then little information is lost by ignoring the corresponding principal components. For large samples, $\widehat{\delta}_q$ is approximately normally distributed:

$$\widehat{\delta}_q \dot{\sim} \text{N}\left[\delta_q, \frac{\left[2\delta_q^2 \sum_{i=1}^{d-q} \lambda_i^2 + 2(1 - \delta_q)^2 \sum_{i=d-q+1}^{d} \lambda_i^2\right]}{n\left(\sum_{i=1}^{d} \lambda_i\right)^2}\right],$$

where $n = N - r$. A one-sided large sample 95% confidence for δ_q is $(0, U)$, where

$$U = \widehat{\delta}_q + 1.645 \frac{\left[2\widehat{\delta}_q^2 \sum_{i=1}^{d-q} \ell_i^2 + 2(1-\widehat{\delta}_q)^2 \sum_{i=d-q+1}^{d} \ell_i^2 \right]^{\frac{1}{2}}}{\sqrt{n} \sum_{i=1}^{d} \ell_i}.$$

Proof: Use the delta method along with the asymptotic results in Theorem 11.6. □

The sample eigenvalues $\ell_1, \ldots, \ell_d$ give information about the dimension of the data. If the last few eigenvalues are very small, then little information is lost by ignoring the corresponding principal components. One model for which the smallest q eigenvalues are equal in magnitude and for which the last q components can be ignored is the following. Let $\mathbf{\Gamma}$ be an orthogonal matrix and partition $\mathbf{\Gamma}$ as $\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_1 & \mathbf{\Gamma}_2 \end{pmatrix}$, where $\mathbf{\Gamma}_1$ has dimension $d \times (d-q)$ and $\mathbf{\Gamma}_2$ has dimension $d \times q$. Also, let $\mathbf{\Lambda}_1 = \text{Diag}(\lambda_1, \ldots, \lambda_{d-q})$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d-q} > 0$. Suppose that $\mathbf{y}$ is a d-vector that can be written as

$$y = \boldsymbol{\mu} + \mathbf{\Gamma}_1 \mathbf{z} + \boldsymbol{\varepsilon},$$

where $\mathbf{z}$ is a $(d-q) \times 1$ random vector with distribution $\text{N}(\mathbf{0}, \mathbf{\Lambda}_1)$, $\boldsymbol{\varepsilon}$ is a random d-vector with distribution $\text{N}(\mathbf{0}, \theta^2 \mathbf{I}_d)$, and $z \perp\!\!\!\perp \boldsymbol{\varepsilon}$. Then,

$$\mathbf{y} \sim \text{N}(\boldsymbol{\mu}, \mathbf{\Sigma}), \text{ where}$$

$$\mathbf{\Sigma} = \mathbf{\Gamma}_1 \mathbf{\Lambda}_1 \mathbf{\Gamma}_1' + \theta^2 \mathbf{I}_d = \mathbf{\Gamma}_1 \mathbf{\Lambda}_1 \mathbf{\Gamma}_1' + \theta^2 \mathbf{\Gamma}\mathbf{\Gamma}' = \mathbf{\Gamma} \begin{pmatrix} \mathbf{\Lambda}_1 + \theta^2 \mathbf{I}_{q-d} & \mathbf{0} \\ \mathbf{0} & \theta^2 \mathbf{I}_q \end{pmatrix} \mathbf{\Gamma}'.$$

Note that the eigenvalues of $\mathbf{\Sigma}$ are $\lambda_1 + \theta^2, \lambda_2 + \theta^2, \ldots, \lambda_{d-q} + \theta^2, \theta^2, \ldots, \theta^2$. The smallest q eigenvalues of $\mathbf{\Sigma}$ are equal with common value θ^2. If such a model is suspected to be true, then the investigators might wish to test the hypothesis $\text{H}_0 \colon \lambda_{d-q+1} = \lambda_{d-q+2} = \cdots = \lambda_d$. That is, a test that the smallest q eigenvalues are equal.

Theorem 11.8 (Test of Partial Sphericity) *The LR test rejects $\text{H}_0 \colon \lambda_{d-q+1} = \lambda_{d-q+2} = \cdots = \lambda_d$ for large Q, where*

$$Q = -n \sum_{j=d-q+1}^{d} \ln \left(\frac{\ell_j}{\bar{\ell}_q} \right), \text{ and } \bar{\ell}_q = \frac{1}{q} \sum_{j=d-q+1}^{d} \ell_j.$$

The asymptotic null distribution of Q is χ_f^2, where $f = \frac{1}{2}(q-1)(q+2)$. The χ^2 approximation is can be improved by using

$$Q^* = -\left[n - d - \frac{2q^2 + q + 2}{6q} + \sum_{j=1}^{d-q} \left(\frac{\bar{\ell}_q}{\ell_j - \bar{\ell}_q} \right)^2 \right] \sum_{j=d-q+1}^{d} \ln \left(\frac{\ell_j}{\bar{\ell}_q} \right),$$

where $n = N - r$ rather than Q as the test statistic. □

The multiplier on Q^* in Theorem 11.8 incorporates a Bartlett correction. The Bartlett correction was proposed by D. N. Lawley (Tests of significance of the latent roots of covariance and correlation matrices, *Biometrika*, 1956, **43**, 128–136) and confirmed by A. T. James [Tests of equality of the latent roots of the covariance matrix, In P. R. Krishnaiah (Ed.) *Multivariate Analysis*, Vol II, pp. 205–218, New York: Academic Press]. MATLAB has an m file (barttest) which performs the test. The MATLAB program does not use the Bartlett correction.

11.3 INFERENCE ON PRINCIPAL COMPONENTS UNDER NON-NORMALITY

11.3.1 References

Boik, R. J. (1998). A local parameterization of orthogonal and semi-orthogonal matrices with applications. *Journal of Multivariate Analysis*, **67**, 244–276.

Boik, R. J. (2003). Principal component models for correlation matrices. *Biometrika*, **90**, 679–701.

11.3.2 Asymptotic Distributions

Consider the conventional linear model for the $N \times d$ matrix $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \text{ where } \mathrm{E}(\mathbf{U}) = \mathbf{0} \text{ and } \mathrm{Disp}(\mathbf{U}) = \mathbf{\Sigma} \otimes \mathbf{I}_N.$$

It is assumed in this section that the rows of $\mathbf{U}$ are independently and identically distributed. That is, $\mathbf{u}_i \overset{\text{iid}}{\sim} (\mathbf{0}, \mathbf{\Sigma})$ for $i = 1, \ldots, N$, where $\mathbf{u}_i$ has dimension $d \times 1$ and $\mathbf{u}_i'$ is the i^{th} row of $\mathbf{U}$.

Let $\mathbf{S}$ be the usual unbiased estimator of $\mathbf{\Sigma}$. That is,

$$\mathbf{S} = n^{-1}\mathbf{Y}'\mathbf{AY}, \text{ where } \mathbf{A} = \mathbf{I}_N - \mathbf{H}_x, \quad \mathbf{H}_x = \mathrm{ppo}(\mathbf{X}), \quad n = N - r,$$

and $r = \mathrm{rank}(\mathbf{X})$. Denote $\mathrm{vec}(\mathbf{S})$ by $\mathbf{s}$ and denote $\mathrm{vec}(\mathbf{\Sigma})$ by $\boldsymbol{\sigma}$. Under mild regularity conditions, it follows from the central limit theorem that the asymptotic distribution of $\sqrt{n}(\mathbf{s} - \boldsymbol{\sigma})$ is multivariate normal. This result is summarized in Theorem 11.9.

Theorem 11.9 (Asymptotic Distribution of S) *If the required moments exist, then*

$$\sqrt{n}(\mathbf{s} - \boldsymbol{\sigma}) \xrightarrow{\text{dist}} \mathrm{N}(\mathbf{0}, \mathbf{\Omega}_\infty), \text{ where } \mathbf{\Omega}_\infty = \lim_{N \to \infty} \mathbf{\Omega}_n \text{ and } \mathbf{\Omega}_n = \mathrm{Var}\left[\sqrt{n}(\mathbf{s} - \boldsymbol{\sigma})\right].$$

Furthermore, if $\mathbf{Y}$ *has a multivariate normal distribution, then*

$$\mathbf{\Omega}_n = \mathbf{\Omega}_\infty = 2\mathbf{N}_d\left(\mathbf{\Sigma} \otimes \mathbf{\Sigma}\right), \text{ where } \mathbf{N}_d = \frac{1}{2}\left(\mathbf{I}_{d^2} + \mathbf{I}_{(d,d)}\right).$$

More generally, Boik (1998) showed that

$$\mathbf{\Omega}_n = 2\mathbf{N}_d\left(\mathbf{\Sigma} \otimes \mathbf{\Sigma}\right) + c_1\left[\mathbf{\Xi} - \boldsymbol{\sigma}\boldsymbol{\sigma}' - 2\mathbf{N}_d\left(\mathbf{\Sigma} \otimes \mathbf{\Sigma}\right)\right], \text{ where}$$

$$\mathbf{\Xi} = \mathrm{E}\left(\mathbf{u}_i\mathbf{u}_i' \otimes \mathbf{u}_i\mathbf{u}_i'\right), \quad c_1 = \frac{1}{n}\,\mathrm{tr}\left(\mathbf{A}^{\odot 2}\right),$$

and $\odot$ *is the element-wise operator. That is,*

$$\mathbf{A}^{\odot 2} = \begin{pmatrix} a_{11}^2 & a_{12}^2 & \cdots & a_{1d}^2 \\ a_{21}^2 & a_{22}^2 & \cdots & a_{2d}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{d1}^2 & a_{d2}^2 & \cdots & a_{dd}^2 \end{pmatrix}.$$

Boik (1998) also obtained an unbiased estimator of $\mathbf{\Omega}_n$*. The estimator is*

$$\widehat{\mathbf{\Omega}}_n = \frac{n}{n-1}2\mathbf{N}_d\left(\mathbf{S} \otimes \mathbf{S}\right) + a_1\widehat{\mathbf{\Xi}} - a_2\left[\mathbf{ss}' + 2\mathbf{N}_d\left(\mathbf{S} \otimes \mathbf{S}\right)\right], \text{ where}$$

$$\widehat{\mathbf{\Xi}} = \frac{1}{n}\sum_{i=1}^{N}(\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i' \otimes \widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i') = \frac{1}{n}(\mathbf{Y}'\mathbf{A} \otimes \mathbf{Y}'\mathbf{A})\left[\sum_{i=1}^{N}(\mathbf{e}_i\mathbf{e}_i' \otimes \mathbf{e}_i\mathbf{e}_i')\right](\mathbf{AY} \otimes \mathbf{AY}),$$

$$a_1 = \frac{nc_1}{(n+2)c_2 - 3c_1^2}, \quad a_2 = \frac{n\left[2c_2 + (n-3)c_1^2\right]}{(n-1)\left[(n+2)c_2 - 3c_1^2\right]}, \quad c_2 = \frac{1}{n}\mathbf{1}_N'\mathbf{A}^{\odot 4}\mathbf{1}_N,$$

$\mathbf{e}_i$ *is the* i^{th} *column of* $\mathbf{I}_N$*,* $\widetilde{\mathbf{u}}_i'$ *is the* i^{th} *row of the residual matrix* $\widetilde{\mathbf{U}} = (\mathbf{I}_N - \mathbf{H}_x)\mathbf{Y} = \mathbf{AY}$*. That is,* $\mathbf{u}_i = \mathbf{Y}'\mathbf{Ae}_i$*. If* $\mathbf{X} = \mathbf{1}_N$*, then* a_1 *and* a_2 *simplify to*

$$a_1 = \frac{n^2}{(n-1)(n-2)} \text{ and } a_2 = \frac{n(n^2-2)}{(n+1)(n-1)(n-2)}, \text{where } n = N - 1.$$

□

Suppose that the eigenvalues of $\boldsymbol{\Sigma}$ are sorted in decreasing order from largest to smallest. Furthermore, suppose that the number of distinct eigenvalues is only k rather than d. Denote the i^{th} largest distinct eigenvalue by φ_i, denote the multiplicity of φ_i by m_i, and let $\mathbf{m}$ be the k-vector of multiplicities. Then,

$$\boldsymbol{\lambda} = \mathbf{T}\boldsymbol{\varphi}, \text{ where } \mathbf{T} = \bigoplus_{j=1}^{k} \mathbf{1}_{m_j} \text{ and } \boldsymbol{\varphi} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_k \end{pmatrix}.$$

Furthermore, $\text{vec}(\boldsymbol{\Lambda})$ can be written as

$$\text{vec}(\boldsymbol{\Lambda}) = \mathbf{L}_d \mathbf{T}\boldsymbol{\varphi}, \text{ where } \mathbf{L}_d = \sum_{i=1}^{d} (\mathbf{e}_i \otimes \mathbf{e}_i)\mathbf{e}_i',$$

and $\mathbf{e}_i$ is the i^{th} column of $\mathbf{I}_d$.

Let $\widehat{\boldsymbol{\varphi}}$ be the estimator of $\boldsymbol{\varphi}$ given by

$$\widehat{\boldsymbol{\varphi}} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\ell},$$

where $\boldsymbol{\ell}$ is the vector of sample eigenvalues. It can be shown that, under normality, $\widehat{\boldsymbol{\varphi}}$ is the MLE of $\boldsymbol{\varphi}$. The asymptotic distribution of $\widehat{\boldsymbol{\varphi}}$ under general conditions was obtained by Boik (1998). The result is summarized in Theorem 11.10.

Theorem 11.10 (Distribution of Sample Eigenvalues) . *Write the covariance in diagonal form as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'$, where the diagonal entries in $\boldsymbol{\Lambda}$ are sorted from largest to smallest. Then, under mild regularity conditions, the asymptotic distribution of $\widehat{\boldsymbol{\varphi}}$ is*

$$\sqrt{n}(\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow{\text{dist}} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi},\infty}), \text{ where } \boldsymbol{\Sigma}_{\boldsymbol{\varphi},\infty} = \lim_{N\to\infty} \boldsymbol{\Sigma}_{\boldsymbol{\varphi},n},$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\varphi},n} = \mathbf{D}_m^{-1}\mathbf{T}'\mathbf{L}_d'(\boldsymbol{\Gamma}' \otimes \boldsymbol{\Gamma}')\boldsymbol{\Omega}_n(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{L}_d\mathbf{T}\mathbf{D}_m^{-1}, \text{ and } \mathbf{D}_m = \text{Diag}(m_1, m_2, \ldots, m_k) = \mathbf{T}'\mathbf{T}.$$

□

Quantities such as δ_q, the ratio of the sum of the smallest q eigenvalues to the sum of all eigenvalues can be expressed as

$$\delta_{c,h} = \frac{\mathbf{c}'\boldsymbol{\varphi}}{\mathbf{h}'\boldsymbol{\varphi}},$$

where $\mathbf{c}$ and $\mathbf{h}$ are k-vectors of constants. For example, if $d = 6$, $k = 4$, and the vector of multiplicities is $\mathbf{m} = \begin{pmatrix} 1 & 1 & 1 & 3 \end{pmatrix}'$, then the ratio of the sum of the smallest 4 eigenvalues to the sum of all eigenvalues is $\delta_{c,h}$, where $\mathbf{c} = \begin{pmatrix} 0 & 0 & 1 & 3 \end{pmatrix}'$ and $\mathbf{h} = \begin{pmatrix} 1 & 1 & 1 & 3 \end{pmatrix}'$. The sample estimator of $\delta_{c,h}$ is

$$\widehat{\delta}_{c,h} = \frac{\mathbf{c}'\widehat{\boldsymbol{\varphi}}}{\mathbf{h}'\widetilde{\boldsymbol{\varphi}}}.$$

The asymptotic distribution is given in Theorem 11.11.

Theorem 11.11 . *Under mild regularity conditions,*

$$\sqrt{n}(\widehat{\delta}_{c,h} - \delta_{c,h}) \xrightarrow{\text{dist}} \text{N}\left[\mathbf{0}, (\mathbf{h}'\boldsymbol{\varphi})^{-2}\mathbf{c}'(\mathbf{I}_k - \mathbf{P}_h)'\boldsymbol{\Sigma}_{\boldsymbol{\varphi},\infty}(\mathbf{I}_k - \mathbf{P}_h)\mathbf{c}\right], \text{ where}$$

$$\mathbf{P}_h = \mathbf{h}(\mathbf{h}'\boldsymbol{\varphi})^{-1}\boldsymbol{\varphi}'.$$

Note that $\mathbf{P}_h$ is the projection operator that projects onto $\mathcal{R}(\mathbf{h})$ along $\mathcal{N}(\boldsymbol{\varphi}')$.

Large sample confidence intervals for $\delta_{c,h}$ can be obtained by substituting $\widehat{\boldsymbol{\varphi}}$ for $\boldsymbol{\varphi}$ in the variance term in Theorem 11.11, and then inverting the usual pivotal quantity based on

$$\widehat{\delta}_{c,h} \dot{\sim} \text{N}\left(\delta_{c,h}, \sigma_\delta^2\right), \text{ where } \sigma_\delta^2 = \frac{\mathbf{c}'(\mathbf{I}_k - \mathbf{P}_h)'\boldsymbol{\Sigma}_{\boldsymbol{\varphi},n}(\mathbf{I}_k - \mathbf{P}_h)\mathbf{c}}{n(\mathbf{h}'\boldsymbol{\varphi})^2},$$

$$\text{and } \sigma_\delta^2 \text{ is estimated by } \widehat{\sigma}_\delta^2 = \frac{\mathbf{c}'(\mathbf{I}_k - \widehat{\mathbf{P}}_h)'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varphi},n}(\mathbf{I}_k - \widehat{\mathbf{P}}_h)\mathbf{c}}{n(\mathbf{h}'\widehat{\boldsymbol{\varphi}})^2}.$$

11.4 PRINCIPAL COMPONENT SCORES

Consider the model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U},$$

where $\mathbf{Y}$ is an $N \times d$ random matrix with dispersion $(\mathbf{\Sigma} \otimes \mathbf{I}_N)$; $\mathbf{X}$ is a known $N \times p$ design matrix with rank r; and $\mathbf{B}$ is a $p \times d$ matrix of unknown regression coefficients. The usual unbiased estimator of $\mathbf{\Sigma}$ is

$$\mathbf{S} = \frac{1}{n}\mathbf{Y}'(\mathbf{I}_N - \mathbf{H}_x)\mathbf{Y},$$

where $\mathbf{H}_x = \text{ppo}(\mathbf{X})$ and $n = N - r$. It is of interest to compute an $N \times d$ matrix, $\mathbf{Z}$, of PC scores (i.e., principal components)

11.4.1 Raw PC Scores

Let

$$\mathbf{\Sigma} = \mathbf{\Gamma\Lambda\Gamma}'$$

be the diagonal form of $\mathbf{\Sigma}$, where the diagonal entries in $\mathbf{\Lambda}$ are ordered from largest to smallest; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. The usual estimator of these quantities is obtained by computing the diagonal form of $\mathbf{S}$:

$$\mathbf{S} = \mathbf{GLG}',$$

where $\mathbf{L}$ is the diagonal matrix containing the ordered sample eigenvalues of $\mathbf{S}$ and $\mathbf{G}$ is the corresponding matrix of sample eigenvectors of $\mathbf{S}$.

There are several sets of PC scores that one could compute. Here are two possibilities:

$$\begin{aligned} \mathbf{Z}_1 &= (\mathbf{Y} - \mathbf{XB})\mathbf{\Gamma} \text{ and} \\ \mathbf{Z}_2 &= (\mathbf{Y} - \mathbf{XB})\mathbf{\Gamma\Lambda}^{-\frac{1}{2}}. \end{aligned}$$

The moments of these principal components are the following:

$$\begin{aligned} \text{E}(\mathbf{Z}_1) &= \mathbf{0}; \quad \text{Disp}(\mathbf{Z}_1) = (\mathbf{\Lambda} \otimes \mathbf{I}_N); \quad \text{Cov}[\text{vec}(\mathbf{Y}), \text{vec}(\mathbf{Z}_1)] = (\mathbf{\Gamma\Lambda} \otimes \mathbf{I}_N) \\ \text{E}(\mathbf{Z}_2) &= \mathbf{0}; \quad \text{Disp}(\mathbf{Z}_2) = \mathbf{I}_{Nd}; \text{ and } \text{Cov}[\text{vec}(\mathbf{Y}), \text{vec}(\mathbf{Z}_2)] = (\mathbf{\Gamma\Lambda}^{\frac{1}{2}} \otimes \mathbf{I}_N). \end{aligned}$$

The second set of scores, $\mathbf{Z}_2$, contains the same information as the first set, $\mathbf{Z}_1$. The scaling of $\mathbf{Z}_2$ is performed so that scores are equally variable.

The usual predictors of these scores are obtained by substituting estimators for the unknown quantities:

$$\begin{aligned} \widehat{\mathbf{Z}}_1 &= (\mathbf{I} - \mathbf{H}_x)\mathbf{YG} \text{ and} \\ \widehat{\mathbf{Z}}_2 &= (\mathbf{I} - \mathbf{H}_x)\mathbf{YGL}^{-\frac{1}{2}}. \end{aligned}$$

11.4.2 Standardized PC Scores

Often, PCA is performed on the correlation matrix rather than on the covariance matrix. The population correlation matrix is

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{D}^{-\frac{1}{2}},$$

where $\mathbf{D} = \text{Diag}(\mathbf{\Sigma})$; i.e., a diagonal matrix containing the variances of the d variables on the diagonal. The usual estimator of $\mathbf{R}$ is

$$\widehat{\mathbf{R}} = \widehat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{S}\widehat{\mathbf{D}}^{-\frac{1}{2}},$$

where $\mathbf{D} = \text{Diag}(\mathbf{S})$.

Let

$$\mathbf{R} = \mathbf{\Gamma\Lambda\Gamma}'$$

be the diagonal form of $\mathbf{R}$, where the diagonal entries in $\mathbf{\Lambda}$ are ordered from largest to smallest; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. The usual estimator of these quantities is given by the diagonal form of the sample correlation matrix:

$$\widehat{\mathbf{R}} = \mathbf{GLG}'.$$

Two sets of PC scores, analogous to $\mathbf{Z}_1$ and $\mathbf{Z}_2$, can be computed:

$$\begin{aligned} \mathbf{Z}_3 &= \mathbf{F\Gamma} \text{ and} \\ \mathbf{Z}_4 &= \mathbf{F\Gamma\Lambda}^{-\frac{1}{2}}, \end{aligned}$$

where $\mathbf{F} = (\mathbf{Y} - \mathbf{XB})\mathbf{D}^{-\frac{1}{2}}$. Note that $\text{Disp}(\mathbf{F}) = (\mathbf{R} \otimes \mathbf{I}_n)$. The moments of these principal components are the following:

$$\begin{aligned} \text{E}(\mathbf{Z}_3) &= \mathbf{0}; \quad \text{Disp}(\mathbf{Z}_3) = (\mathbf{\Lambda} \otimes \mathbf{I}_N); \quad \text{Cov}[\text{vec}(\mathbf{F}), \text{vec}(\mathbf{Z}_3)] = (\mathbf{\Gamma\Lambda} \otimes \mathbf{I}_N) \\ \text{E}(\mathbf{Z}_4) &= \mathbf{0}; \quad \text{Disp}(\mathbf{Z}_4) = \mathbf{I}_{Nd}; \text{ and } \text{Cov}[\text{vec}(\mathbf{F}), \text{vec}(\mathbf{Z}_4)] = (\mathbf{\Gamma\Lambda}^{\frac{1}{2}} \otimes \mathbf{I}_N). \end{aligned}$$

Caution, $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ refer to eigenvectors and -values of the correlation matrix. These are not the same quantities that appear in the moments of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Also, note that $\text{Corr}[\text{vec}(\mathbf{F}), \text{vec}(\mathbf{Z}_4)] = (\mathbf{\Gamma\Lambda}^{\frac{1}{2}} \otimes \mathbf{I}_N)$. The matrix $\mathbf{\Gamma\Lambda}^{\frac{1}{2}}$ represents the correlations between the standardized data (rows of $\mathbf{F}$) and the PC scores. It is sometimes called the Factor Loading Matrix. Rencher (2002) cautions against interpreting PCs by using the Factor Loading matrix. The second set of scores, $\mathbf{Z}_4$, contains the same information as the first set, $\mathbf{Z}_3$.

The usual predictors of these scores are obtained by substituting estimators for the unknown quantities:

$$\begin{aligned} \widehat{\mathbf{Z}}_3 &= (\mathbf{I} - \mathbf{H}_x)\mathbf{Y}\widehat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{G} \text{ and} \\ \widehat{\mathbf{Z}}_4 &= (\mathbf{I} - \mathbf{H}_x)\mathbf{Y}\widehat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{GL}^{-\frac{1}{2}}. \end{aligned}$$

11.5 COMMON PRINCIPAL COMPONENTS AND GENERALIZATIONS

11.5.1 References

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, **89**, 159–182.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models.* New York: John Wiley & Sons.

11.5.2 The CPC Model

Suppose that a sample of size N_i is obtained from each of k populations. On each case, a d dimensional random vector is observed. One model for this data is

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \text{ where } \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_k \end{pmatrix}, \quad \text{Disp}(\mathbf{Y}_i) = \mathbf{\Sigma_i} \otimes \mathbf{I}_N,$$

and $\mathbf{Y}_i$ is the $N_i \times d$ matrix of observations from the i^{th} population. Flury's common principal components (CPC) model for the heterogeneous covariance matrices is

$$\mathbf{\Sigma}_i = \mathbf{\Gamma\Lambda}_i\mathbf{\Gamma}' \text{ for } i = 1, \ldots, k.$$

By modeling the covariance matrices, the number of parameters has been reduced from $kd(d+1)/2$ to $d(d-1)/2+dk$. The model is called common principal components because the coefficients for the principal components (i.e., the columns of $\mathbf{\Gamma}$) are common across populations.

As an example, consider the following parameters from three populations:

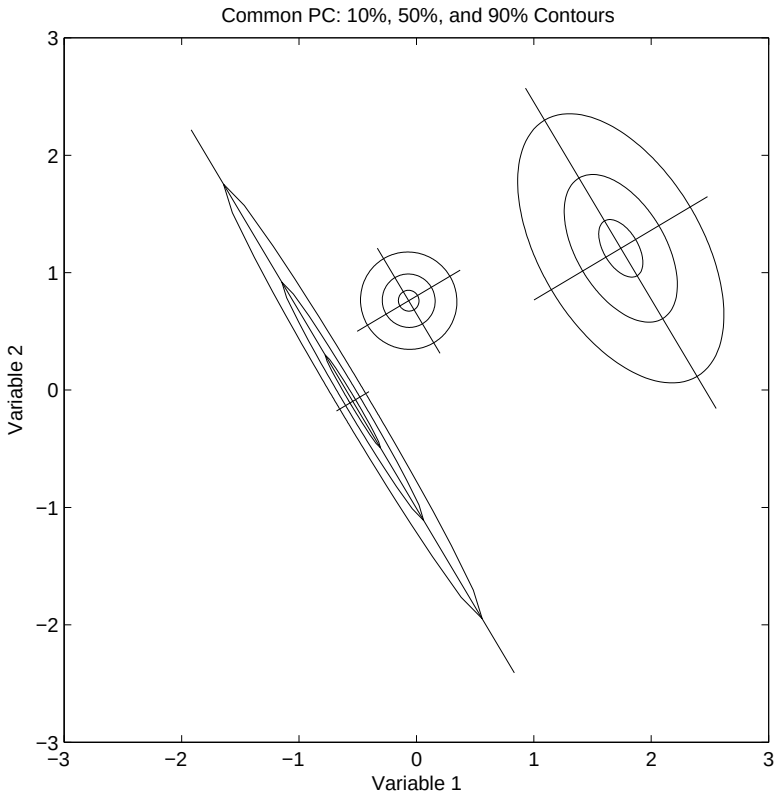
$$\mathbf{\Gamma} = \begin{pmatrix} -0.5115 & 0.8593 \\ 0.8593 & 0.5115 \end{pmatrix},$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1.7409 \\ 1.2074 \end{pmatrix}, \quad \boldsymbol{\lambda}_1 = \begin{pmatrix} 0.3503 \\ 0.1029 \end{pmatrix} \Longrightarrow \mathbf{\Sigma}_1 = \begin{pmatrix} 0.1676 & -0.1087 \\ -0.1087 & 0.2855 \end{pmatrix},$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} -0.0658 \\ 0.7613 \end{pmatrix}, \quad \boldsymbol{\lambda}_2 = \begin{pmatrix} 0.0378 \\ 0.0362 \end{pmatrix} \Longrightarrow \mathbf{\Sigma}_2 = \begin{pmatrix} 0.0366 & -0.0007 \\ -0.0007 & 0.0373 \end{pmatrix},$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} -0.5418 \\ -0.0965 \end{pmatrix}, \text{ and } \boldsymbol{\lambda}_3 = \begin{pmatrix} 1.0068 \\ 0.0036 \end{pmatrix} \Longrightarrow \mathbf{\Sigma}_3 = \begin{pmatrix} 0.2660 & -0.4409 \\ -0.4409 & 0.7443 \end{pmatrix}.$$

The covariance matrices are plotted below.



Note that the three ellipses are aligned. This occurs because the three covariance matrices share the same eigenvectors. It is the eigenvectors that determine the orientation of the plots. The magnitude of the major and minor axes differ among the three ellipses. This occurs because the eigenvalues differ among the three covariance matrices. The eigenvalues determine the shape of the ellipses.

Flury derived an algorithm for computing MLEs of the parameters under normality. Also, he derived the asymptotic distribution of the MLEs and a likelihood ratio test of $\mathrm{H}_1\colon \mathbf{\Sigma}_i = \mathbf{\Gamma}\mathbf{\Lambda}_i\mathbf{\Gamma}'$ against $\mathrm{H}_a\colon \mathbf{\Sigma}_i > 0$.

In partial common principal components, only q of the d components are common. The remainder are population-specific. In the common space model, all components are population-specific, but q of the d components in each population share the same eigenspace.

11.5.3 Extensions of the CPC Model

Boik (2002) extended Flury's model in the following manner:

1. The eigenvalues of the k covariance matrices are modeled to allow for arbitrary multiplicities and to allow relationships among eigenvalues from the k populations.

2. The eigenvectors from any set of the k populations are allowed to (a) be distinct and functionally independent, (b) share the same eigenspace, or (c) be identical. The eigenvalues that correspond to these sets can be ordered or unordered.

3. Second-order asymptotic distributions were derived for all parameter estimators under normality and first-order distributions were derived under nonnormality.

4. Bartlett corrections were derived for performing model comparison tests using the likelihood ratio test statistic.

5. The asymptotic null distribution of the likelihood ratio test statistic was obtained without assuming normality.

11.6 HOTELLING'S POWER ALGORITHM

Suppose we desire the maximum root and associated vector of a psd matrix $\mathbf{A}$. Write $\mathbf{A}$ in diagonal form as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$.

Theorem 11.12 *Randomly choose a $d \times 1$ vector and denote the vector by $\mathbf{t}_0$. Define $\mathbf{t}_{i+1}$*

$$\mathbf{t}_{i+1} = \frac{\mathbf{A}\mathbf{t}_i}{\sqrt{\mathbf{t}_i'\mathbf{A}^2\mathbf{t}_i}}.$$

Then as $i \to \infty$, $\mathbf{t}_{i+1}$ converges to $\mathbf{u}_1$ provided that $\lambda_1 > \lambda_2$.
Proof: HW.

□

To obtain the j^{th} largest component, the above iterative method is used but $\mathbf{A}$ is replaced by

$$\mathbf{A}_{j-1} = \mathbf{A} - \sum_{i=1}^{j-1} \lambda_i \mathbf{u}_i \mathbf{u}_i'.$$

11.7 SINGULAR VALUE DECOMPOSITION

Let $\mathbf{Y}$ be any real $a \times b$ matrix of rank $r \leq \min(a, b)$. Eckart and Young (1936, *Psychometrika*, *1*, 211–218) showed that $\mathbf{Y}$ can be written as $\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$ where $\mathbf{U}$ is $a \times r$, $\mathbf{U}'\mathbf{U} = \mathbf{I}_r$, $\mathbf{V}$ is $b \times r$, $\mathbf{V}'\mathbf{V} = \mathbf{I}_r$, $\mathbf{\Lambda}$ is an $r \times r$ diagonal matrix having positive entries on the diagonal, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$. If the λ_i's are distinct, the decomposition can be made to be unique by imposing identifiability restriction on $\mathbf{U}$ or $\mathbf{V}$. For example, one suitable set of restrictions requires that the first non-zero entry in each column of $\mathbf{U}$ be positive. The expression $\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$ is called the singular value decomposition (SVD) of $\mathbf{Y}$. The λ_i's are called the singular values.

The SVD of $\mathbf{Y}$ can be obtained as follows. The columns of $\mathbf{U}$ are the ch. vectors of $\mathbf{Y}\mathbf{Y}'$. The columns of $\mathbf{V}$ are the ch. vectors of $\mathbf{Y}'\mathbf{Y}$. The nonzero roots of $\mathbf{Y}\mathbf{Y}'$ and $\mathbf{Y}'\mathbf{Y}$ are the squares of the λ_i's. That is

$$\mathbf{Y}\mathbf{Y}' = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}' \quad \text{and} \quad \mathbf{Y}'\mathbf{Y} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}'.$$

The rank-m matrix, for $m \leq r$, which minimizes $\|\mathbf{Y} - \mathbf{M}\|^2$ is given by

$$\mathbf{M} = \sum_{i=1}^{m} \lambda_i \mathbf{u}_i \mathbf{v}_i'.$$

The Moore-Penrose inverse of $\mathbf{Y}$, say $\mathbf{Y}^+$ is $\mathbf{Y}^+ = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}'$.

11.8 BIPLOTS

If $\mathbf{Y}$ can be approximated fairly well by a rank two matrix, then the matrix can be plotted in two-dimensional space. Suppose $\mathbf{Y} \approx \mathbf{R}\mathbf{C}'$ for $\mathbf{R} : a \times 2$ and $\mathbf{C} : b \times 2$. Then the rows of $\mathbf{R}$ and $\mathbf{C}$ can be plotted in 2-space. Gabriel has suggestions on how to choose $\mathbf{R}$ and $\mathbf{C}$ and how to interpret them.

Chapter 12

FACTOR ANALYSIS

12.1 THE FACTOR ANALYSIS MODEL

The basic FA model can be written as follows:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is a $d \times 1$ random vector; $\boldsymbol{\Gamma}$ is an unknown $d \times m$ matrix of constants; $\mathbf{f}$ is an $m \times 1$ unobserved random vector having mean $\mathbf{0}$ and dispersion $\boldsymbol{\Sigma}_f$; $\boldsymbol{\varepsilon}$ is a $d \times 1$ unobserved random "error" vector having mean $\mathbf{0}$ and diagonal dispersion $\boldsymbol{\Psi}$; and $\text{cov}(\mathbf{f}, \boldsymbol{\varepsilon}) = \mathbf{0}$. The problem is to estimate $\boldsymbol{\Psi}$, $\boldsymbol{\Gamma}$ and sometimes to predict $\mathbf{f}$. Without loss of generality it can be assumed that $\boldsymbol{\Sigma}_f = \mathbf{I}_m$. In this case,

$$\mathbf{y} \sim (\boldsymbol{\mu}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}).$$

The parameter estimation problem is to use the sample covariance matrix, $\mathbf{S}$, to estimate $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$.

The matrix $\boldsymbol{\Gamma}$ is called the factor loading matrix. For $\boldsymbol{\Sigma}_f = \mathbf{I}_m$, it is easy to show that

$$\text{cov}(\mathbf{y}, \mathbf{f}) = \boldsymbol{\Gamma}.$$

12.2 THE PROBLEM OF NON-UNIQUENESS

12.2.1 Maximum Number of Unique Factors

The covariance matrix $\boldsymbol{\Sigma}$ has $d(d+1)/2$ parameters. The factor model $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$ has $md - m(m-1)/2 + d$ parameters. The quantity $m(m-1)/2$ represents the number of entries of $\boldsymbol{\Gamma}$ that can be annihilated by postmultiplying by an orthogonal matrix. That is $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}\mathbf{Q}\mathbf{Q}'\boldsymbol{\Gamma}'$, where $\mathbf{Q}$ is any orthogonal matrix. The orthogonal matrix that annihilates entries in $\boldsymbol{\Gamma}$ can be computed using the QR decomposition:

$$\boldsymbol{\Gamma}' = \mathbf{Q}\mathbf{R},$$

where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper triangular. Thus $\boldsymbol{\Gamma}\mathbf{Q} = \mathbf{R}'$ is lower triangular; the $m(m-1)/2$ entries in the upper right-hand corner of $\boldsymbol{\Gamma}$ have been annihilated. Thus, a necessary condition for uniqueness is that $d(d+1)/2 \geq md - m(m-1)/2 + d$ or, equivalently,

$$m \leq \frac{2d + 1 - \sqrt{8d-1}}{2}.$$

The degrees of freedom remaining after fitting an m factor model are

$$df = \frac{d(d+1)}{2} - md + \frac{m(m-1)}{2} - d = \frac{(d-m)^2 - (d+m)}{2}.$$

Below is a table of maximum values of m for selected values of d.

d	$\max m$
1	0
2	0
3	1
4	1
5	2
6	3
7	3
8	4
9	5
10	6

12.2.2 Rotation Indeterminacy

Let $\mathbf{T}$ be any orthogonal matrix of order m. Let $\mathbf{\Gamma}^* = \mathbf{\Gamma T}$ and $\mathbf{f}^* = \mathbf{T}'\mathbf{f}$. Then

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{\Gamma f} + \boldsymbol{\varepsilon} \Longleftrightarrow \mathbf{y} = \boldsymbol{\mu} + \mathbf{\Gamma}^*\mathbf{f}^* + \boldsymbol{\varepsilon},$$

$$\text{var}(\mathbf{f}^*) = \mathbf{I}, \quad \text{and} \quad \text{var}(\mathbf{y}) = \mathbf{\Gamma}^*\mathbf{\Gamma}^{*\prime} + \mathbf{\Psi} = \mathbf{\Gamma\Gamma}' + \mathbf{\Psi}.$$

The problem of arbitrary orthogonal rotation is dealt with in two ways. During the estimation phase (when ML estimation is used), it is required that $\widehat{\mathbf{\Gamma}}'\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Gamma}}$ be a diagonal matrix. This is equivalent to choosing $\mathbf{T}$ to be the matrix of ch. vectors of $\widehat{\mathbf{\Gamma}}'\mathbf{\Psi}^{-1}\widehat{\mathbf{\Gamma}}$ and replacing $\widehat{\mathbf{\Gamma}}$ by $\widehat{\mathbf{\Gamma}}\mathbf{T}$.

After the estimation phase, an alternative $\mathbf{T}$ can be used. Let $\mathbf{\Gamma}^* = \mathbf{\Gamma T}$ for some $\mathbf{T}$. One strategy is to choose $\mathbf{T}$ so that the factor loading matrix $\mathbf{\Gamma}^*$ has a simple structure. One such structure is when each column of $\mathbf{\Gamma}^*$ has entries which are either near zero or large in absolute value. In this case, the interpretation of the factor is simplified. To find the $\mathbf{T}$ which yields this "simple structure", the varimax criterion often is used. Write $\mathbf{\Gamma}^*$ as $\mathbf{\Gamma}^* = \{\gamma_{ij}^*\}$. Let $g_{ij} = {\gamma_{ij}^*}^2$ and define $\mathbf{G}\colon d \times m$ as $\mathbf{G} = \{g_{ij}\}$. Let $\mathbf{V}\colon m \times m$ be the "covariance" matrix for the columns of $\mathbf{G}$. That is

$$\mathbf{V} = \frac{\mathbf{G}'(\mathbf{I}_d - \mathbf{H}_1)\mathbf{G}}{d-1},$$

where $\mathbf{H}_1 = \text{ppo}(\mathbf{1}_d)$. The varimax criterion chooses $\mathbf{T}$ such that $\text{tr}(\mathbf{V})$ is maximized.

An alternative criteria is quartimax in which the variance within the rows of $\mathbf{G}$ is maximized. That is, the orthogonal matrix $\mathbf{T}$ is chosen to maximize $\text{tr}(\mathbf{V}^*)$, where

$$\mathbf{V}^* = \frac{\mathbf{G}(\mathbf{I}_m - \mathbf{H}_1)\mathbf{G}'}{m-1},$$

where $\mathbf{H}_1 = \text{ppo}(\mathbf{1}_m)$. It can be shown that $\mathbf{G1}_m$ does not depend on $\mathbf{T}$ so $\mathbf{T}$ can be chosen to maximize $\text{tr}(\mathbf{GG}')$.

12.3 PRINCIPAL COMPONENTS VERSUS FACTOR ANALYSIS

The two procedures, PCA and FA, are often confused. To see how they differ, write $\mathbf{\Sigma}$ in diagonal form as $\mathbf{\Sigma} = \mathbf{\Gamma\Lambda\Gamma}'$ where $\mathbf{\Lambda}$ has the ordered roots on the diagonal. Partition $\mathbf{\Gamma}$ as $\mathbf{\Gamma} = \begin{pmatrix}\mathbf{\Gamma}_1 & \mathbf{\Gamma}_2\end{pmatrix}$, where $\mathbf{\Gamma}_1$ has dimension $d \times m$. Let $\mathbf{\Lambda}_1\colon m \times m$ be the upper left-hand corner matrix of $\mathbf{\Lambda}$ and let $\mathbf{\Lambda}_2$ be the lower right-hand corner matrix of λ.

Using principal components, the best m-dimensional approximation to $\mathbf{y}$ is

$$\widehat{\mathbf{y}} = \boldsymbol{\mu} + \mathbf{A}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\mathbf{A} = \mathbf{\Gamma}_1\mathbf{\Gamma}_1'$. The associated model is

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{\Gamma}_1\mathbf{z}_1 + \boldsymbol{\varepsilon},$$

where $\mathbf{z}_1 = \mathbf{\Gamma}_1'(\mathbf{y} - \boldsymbol{\mu})$. An equivalent model can be obtained by letting $\mathbf{\Gamma}_1^* = \mathbf{\Gamma}_1\mathbf{\Lambda}_1^{\frac{1}{2}}$ and $\mathbf{f} = \mathbf{\Lambda}_1^{-\frac{1}{2}}\mathbf{z}_1$ so that

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{\Gamma}_1^*\mathbf{f} + \boldsymbol{\varepsilon}.$$

This looks much like the FA model. As in the FA model, the factors have an identity covariance matrix and are uncorrelated with the residuals:

$$\text{var}(\mathbf{f}) = \text{var}[\mathbf{\Lambda}_1^{-\frac{1}{2}}\mathbf{\Gamma}_1'(\mathbf{y}-\boldsymbol{\mu})] = \mathbf{I}_m,$$

and

$$\text{cov}(\mathbf{f},\boldsymbol{\varepsilon}) = \text{cov}[\mathbf{\Lambda}_1^{-\frac{1}{2}}\mathbf{\Gamma}_1'(\mathbf{y}-\boldsymbol{\mu}), \mathbf{\Gamma}_2\mathbf{\Gamma}_2'(\mathbf{y}-\boldsymbol{\mu})] = \mathbf{\Lambda}_1^{-\frac{1}{2}}\mathbf{\Gamma}_1'\mathbf{\Sigma}\mathbf{\Gamma}_2\mathbf{\Gamma}_2' = \mathbf{0}.$$

Note, however, that the covariance matrix for the errors is not diagonal:

$$\text{var}(\boldsymbol{\varepsilon}) = \mathbf{\Gamma}_2\mathbf{\Lambda}_2\mathbf{\Gamma}_2'.$$

Thus, the "factors" obtained from principal components do not explain the entire covariance structure.

12.4 MAXIMUM LIKELIHOOD ESTIMATION

Suppose that $\mathbf{Y}$ is a random $N \times d$ matrix that follows the linear model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \text{ where } \text{vec}(\mathbf{U}) \sim \text{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{I}_N),$$

$\mathbf{X}$ is an $N \times p$ matrix of constants, and $\text{rank}(\mathbf{X}) = r$. Denote the usual estimator of $\mathbf{\Sigma}$ by $\mathbf{S}$. That is,

$$\mathbf{S} = \frac{1}{n}\mathbf{Y}'(\mathbf{I}_N - \mathbf{H}_x)\mathbf{Y}, \text{ where } n = N - r \text{ and } \mathbf{H}_x = \text{ppo}(\mathbf{X}).$$

If the factor model holds, then

$$n\mathbf{S} \sim \text{W}_d(n, \mathbf{\Sigma}), \text{ where } \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}.$$

The MLE's of $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ are those that maximize

$$L(\mathbf{\Gamma}, \mathbf{\Psi}|\mathbf{S}) = -\frac{n}{2}\ln|\mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}| - \frac{n}{2}\text{tr}[\mathbf{S}(\mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi})^{-1}].$$

Setting the derivatives to zero yields the equations

$$\mathbf{S}\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}'\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Gamma}} + \mathbf{I})$$

and

$$\widehat{\mathbf{\Psi}} = \text{diag}(\mathbf{S} - \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Gamma}}').$$

To obtain a unique solution, the restriction that $\widehat{\mathbf{\Gamma}}'\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Gamma}}$ is diagonal is imposed.

One appealing property of the MLE approach is that the solution is equivariant with respect to the scale employed. Equivariant estimators satisfy the following property — if $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{\Psi}}$ are mles based on the sample covariance matrix and if $\mathbf{V}$ is a diagonal matrix, then the mles based on $\mathbf{VSV}$ are $\widehat{\mathbf{\Gamma}}^* = \mathbf{V}\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{\Psi}}^* = \mathbf{V}\widehat{\mathbf{\Psi}}\mathbf{V}$. In particular, the MLEs based on a covariance and the MLEs based on a correlation matrix are simple functions of one another. That the MLEs are equivariant with respect to scale can be seen by examining the LR criterion. Suppose that $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{\Psi}}$ maximize $L(\mathbf{\Gamma}, \mathbf{\Psi}|\mathbf{S})$ and $\widehat{\mathbf{\Gamma}}'\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Gamma}}$ is diagonal. Let $\mathbf{R} = \mathbf{DSD}$ where $\mathbf{D}$ is a positive definite diagonal matrix. If $\mathbf{D}$ is defined by $\mathbf{D} = [\text{diag}(\mathbf{S})]^{-\frac{1}{2}}$, then $\mathbf{R}$ is the sample correlation matrix.

Substituting $\mathbf{R}$ for $\mathbf{S}$ in the likelihood function yields

$$L(\mathbf{\Gamma}_r, \mathbf{\Psi}_r|\mathbf{R}) = -\ln|\mathbf{\Gamma}_r\mathbf{\Gamma}_r' + \mathbf{\Psi}_r| - \text{tr}[\mathbf{R}(\mathbf{\Gamma}_r\mathbf{\Gamma}_r' + \mathbf{\Psi}_r)^{-1}]$$

$$= -\ln|\mathbf{D}^{-1}\mathbf{\Gamma}_r\mathbf{\Gamma}_r'\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{\Psi}_r\mathbf{D}^{-1}| - \text{tr}[\mathbf{S}(\mathbf{D}^{-1}\mathbf{\Gamma}_r\mathbf{\Gamma}_r'\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{\Psi}_r\mathbf{D}^{-1})^{-1}]$$

$$-2\ln|\mathbf{D}|$$

$$= -\ln|\mathbf{\Gamma}^*\mathbf{\Gamma}^{*\prime} + \mathbf{\Psi}^*| - \text{tr}[\mathbf{S}(\mathbf{\Gamma}^*\mathbf{\Gamma}^{*\prime} + \mathbf{\Psi}^*)^{-1}] - 2\ln|\mathbf{D}|,$$

where $\mathbf{\Gamma}^* = \mathbf{D}^{-1}\mathbf{\Gamma}_r$ and $\mathbf{\Psi}^* = \mathbf{D}^{-1}\mathbf{\Psi}_r\mathbf{D}^{-1}$. Because $\ln|\mathbf{D}|$ is constant with respect to $\mathbf{\Gamma}$ and $\mathbf{\Psi}$, the likelihood function is maximized by

$$\widehat{\mathbf{\Gamma}}^* = \widehat{\mathbf{\Gamma}} \Longrightarrow \widehat{\mathbf{\Gamma}}_r = \mathbf{D}\widehat{\mathbf{\Gamma}}$$

and

$$\widehat{\mathbf{\Psi}}^* = \widehat{\mathbf{\Psi}} \Longrightarrow \widehat{\mathbf{\Psi}}_r = \mathbf{D}\widehat{\mathbf{\Psi}}\mathbf{D}.$$

Note also that $\widehat{\mathbf{\Gamma}}_r'\widehat{\mathbf{\Psi}}_r^{-1}\widehat{\mathbf{\Gamma}}_r = \widehat{\mathbf{\Gamma}}'\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Gamma}}$ so the diagonal restriction is satisfied.

To test goodness of fit, one can test $\text{H}_0\colon \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}$ against $\text{H}_a\colon \mathbf{\Sigma} > 0$. The degrees of freedom for the LR test are $[(d-m)^2 - (d+m)]/2$. Why?

12.5 PRINCIPAL FACTOR ANALYSIS

The principal factor solution is the minimizer of $\text{tr}[(\mathbf{S}-\boldsymbol{\Sigma})^2]$ subject to $\boldsymbol{\Sigma}=\boldsymbol{\Gamma}\boldsymbol{\Gamma}'+\boldsymbol{\Psi}$. That is, the model is the same as before but a different criterion (least squares loss function) is used. In the i^{th} iteration, the principal factor routine finds the largest m roots and associated vectors of $\mathbf{S}-\widehat{\boldsymbol{\Psi}}_i$. A reasonable initial guess, $\widehat{\boldsymbol{\Psi}}_0$ consists of the inverse of the diagonal elements of $\mathbf{S}^{-1}$.

12.6 ESTIMATING (PREDICTING) FACTOR SCORES

12.6.1 Prediction Approach

If $\mathbf{f}$ were fixed, then the model for $(\mathbf{y}-\boldsymbol{\mu})$ would be

$$(\mathbf{y}-\boldsymbol{\mu}) \sim (\boldsymbol{\Gamma}\mathbf{f}, \boldsymbol{\Psi}).$$

For known $\boldsymbol{\Gamma}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\mu}$ the generalized least squares estimator of $\mathbf{f}$ would be

$$\widehat{\mathbf{f}} = (\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}).$$

In practice, estimates of $\boldsymbol{\Gamma}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\mu}$ would be used.

Of course, $\mathbf{f}$ is not fixed. It is a random vector. Thus, the problem is to predict $\mathbf{f}$ from $\mathbf{y}$ rather than to estimate $\mathbf{f}$. To make such predictions, we need a criterion for judging how good a predictor is. Consider a pair of random vectors $\mathbf{U}$ and $\mathbf{Y}$ having joint density function $f(\mathbf{u},\mathbf{y})$. Given $\mathbf{Y}=\mathbf{y}$, the goal is to predict $\mathbf{u}$. The predictor $\widehat{\mathbf{u}}=h(\mathbf{y})$ is called the best predictor if it minimizes the mean square error of prediction:

$$\begin{aligned} MSE(\widehat{\mathbf{u}}) &= \text{E}[(\mathbf{u}-\widehat{\mathbf{u}})'\mathbf{A}(\mathbf{u}-\widehat{\mathbf{u}})] \\ &= \int\int (\mathbf{u}-\widehat{\mathbf{u}})'\mathbf{A}(\mathbf{u}-\widehat{\mathbf{u}}) f(\mathbf{u},\mathbf{y})\, d\mathbf{u}\, d\mathbf{y}, \end{aligned}$$

where $\mathbf{A}$ is a positive definite matrix (e.g., the inverse of a covariance matrix).

Theorem 12.1 (Best Prediction) *The best predictor of* $\mathbf{U}$ *given that* $\mathbf{Y}=\mathbf{y}$ *has been observed is* $\widehat{\mathbf{u}}=\text{E}(\mathbf{U}|\mathbf{Y}=\mathbf{y})$.
Outline of Proof: Write $f(\mathbf{u},\mathbf{y})$ *as*

$$f(\mathbf{u},\mathbf{y}) = g(\mathbf{u}|\mathbf{y}) \times m(\mathbf{y}),$$

where $g(\mathbf{u}|\mathbf{y})$ *is the conditional density of* $\mathbf{U}$ *given* $\mathbf{Y}=\mathbf{y}$ *and* $m(\mathbf{y})$ *is the marginal density of* $\mathbf{Y}$*. To minimize MSE, consider minimizing MSE for each realization of* $\mathbf{Y}$*. For fixed* $\mathbf{Y}=\mathbf{y}$*, note that* $\widehat{\mathbf{u}}$ *is a constant.*

□

It is easy to show that $\text{E}(\widehat{\mathbf{u}})=\text{E}(\mathbf{U})$ so that the best predictor is unbiased. Note that the joint density $f(\mathbf{u},\mathbf{y})$ need not be normal.

For the FA problem, suppose that

$$(\mathbf{y}-\boldsymbol{\mu})|\mathbf{f},\boldsymbol{\Gamma},\boldsymbol{\Psi} \sim \text{N}(\boldsymbol{\Gamma}\mathbf{f},\boldsymbol{\Psi}),$$

and

$$\mathbf{f} \sim \text{N}(\mathbf{0},\mathbf{I}_m).$$

Then it can be shown (problem 5.17 in Seber) that

$$\mathbf{f}|\mathbf{y},\boldsymbol{\Gamma},\boldsymbol{\Psi},\boldsymbol{\mu} \sim \text{N}\left[(\mathbf{I}+\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}), (\mathbf{I}+\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma})^{-1}\right].$$

The best predictor of $\mathbf{f}$ given that we have observed $\mathbf{y}$ is

$$\widehat{\mathbf{f}} = (\mathbf{I}+\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}).$$

In practice, estimates of $\boldsymbol{\Gamma}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\mu}$ would be used.

12.6.2 Regression Approach

Rencher (2002) describes a regression method for "estimating" factor scores. If $\mathbf{S}$ is replaced by $\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\Gamma}}'+\widehat{\boldsymbol{\Psi}}$, then the regression approach is identical to the prediction approach.

Chapter 13

CLUSTER ANALYSIS

1. Distances: The distance between two objects, a and b, is denoted by $d(a, b)$. The distance must satisfy the following four properties
   (a) $d(a, a) = 0$
   (b) $d(a, b) > 0$ if $a \neq b$
   (c) $d(a, b) = d(b, a)$
   (d) $d(a, b) \leq d(a, c) + d(c, b)$ (triangle inequality)
2. Distance measures in cluster analysis
   (a) Distance between cases
      i. Euclidean distance: $d(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)}$. This is the default in proc cluster.
      ii. Euclidean distance on standardized measures: $d(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{(\mathbf{z}_r - \mathbf{z}_s)'(\mathbf{z}_r - \mathbf{z}_s)}$, where $\mathbf{z}_r = \{z_{ri}\}$ and $z_{ri} = (x_{ri} - \bar{x}_i)/s_i$. This is employed in proc cluster if the standard option is used.
      iii. Mahalanobis distance: $d(\mathbf{x}_r, \mathbf{x}_s) = \sqrt{(\mathbf{x}_r - \mathbf{x}_s)'\mathbf{S}^{-1}(\mathbf{x}_r - \mathbf{x}_s)}$
      iv. Minkowski distance: $d(\mathbf{x}_r, \mathbf{x}_s) = [\sum_{i=1}^{p} |x_{ri} - x_{si}|^m]^{\frac{1}{m}}$
   (b) Distances between variables
      i. $d(x_i, x_j) = \sqrt{1 - r_{ij}}$, where r_{ij} is the correlation between variables i and j
      ii. $d(x_i, x_j) = \sqrt{1 - r_{ij}^2}$, where r_{ij} is the correlation between variables i and j.
3. Hierarchical clustering: divisive techniques
   (a) Description of the approach
      i. Begin with all n cases (or p variables) belonging to one cluster
      ii. Split the cluster into two clusters
      iii. Split one of the two clusters into two clusters
      iv. Continue splitting until each case (or variable) is its own cluster
   (b) Proc cluster does not do divisive clustering
4. Hierarchical clustering: agglomerative techniques
   (a) Description of the approach
      i. Begin with each case (or variable) being its own cluster
      ii. Join the closest two items to form a new cluster. The number of clusters is now $n - 1$ (or $p - 1$).
      iii. Join the closest two clusters to form a new cluster. The number of clusters is now $n - 2$ (or $p - 2$).
      iv. Continue joining clusters until one cluster contains all cases (or variables).
   (b) Proc cluster will do several variations of agglomerative clustering
      i. Method = single. Clusters having nearest neighbors are joined. This method is useful if clusters are not spherical.

ii. Method = complete. Clusters having nearest far neighbors (furthest neighbors) are joined. The far neighbor distance is the maximum distance between an object in the first cluster and an object in the second cluster. This method is useful if clusters are spherical.

iii. Method = average. Clusters having the smallest average distance are joined. This is a compromise between single and complete linkage.

(c) Non-hierarchical clustering

i. K-means clustering. Begin with K seeds (initial cluster centroids). Assign objects to the closest seed. Iterate until no further reassignments are made. Proc fastclus does K means clustering.

ii. Other methods: There are many additional procedures including maximum likelihood mixture model clustering.

(d) Determining the number of clusters

i. Examine the tree structure. Look for natural clusters. Look for breaks in the distance. Look for the point at which the distance is too large.

ii. Cubic clustering criterion. Look for a peak value of CCC that has value 3 or more. (Simulation based criterion).

iii. Examine pseudo F and T^2 statistics. These are tests statistics for testing the hypothesis that the cluster means are identical. These tests do not yield valid p-values because the data are used to form the cluster in the first place. Nonetheless, they do give indices of how far apart the two clusters are that were just joined (T^2) and how far apart the g clusters are from each other F). Small values of T^2 indicate that the two clusters that were just joined are close together. Large values of F indicate that the cluster means of the g clusters are far apart.

Chapter 14

CLASSIFICATION TREES

1. References

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.G. (1984). *Classification and Regression Trees*. Belmont CA: Wadsworth International.

CART. (1998). San Diego CA: Salford Systems.

Clark, L.A., & Pregibon, D. (1992). "Tree-Based Models," in J.M. Chambers and T.J. Hastie (eds) *Statistical Models in S*, pp. 377–420. Pacific Grove, CA: Wadsworth and Brooks-Cole.

De'ath, G., & Fabricius, K.E. (2000). "Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis," *Ecology*, **81**, 3178–3192.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Chapter 7 "Tree-structured Classifiers." Cambridge: Cambridge University Press.

Venables, W.N., & Ripley, B.D. (1999). *Modern Applied Statistics with S-PLUS*, third edition, Chapter 10, "Tree-based methods." New York: Springer-Verlag.

2. Terminology

   (a) Node: possible decision point in the tree

   (b) Root: top node

   (c) Leaf: terminal node

   (d) Subtree of T: a tree whose root is a node of T

   (e) Rooted subtree of T: a subtree whose root is the root of T

   (f) Grow: split a leaf

   (g) Prune: delete one or more subtrees

3. Measures of Leaf Impurity

   (a) Deviance (entropy): Let c be the number of classes and let v be the number of leaves. Condition on the values of the observed variables and denote the conditional probability that an observation is in class k given that it is at leaf i as $p(k|i) = p_{ik}$. Then, the conditional likelihood is

$$L(\mathbf{p}|\mathbf{Y}) = \prod_{i=1}^{v} \prod_{k=1}^{c} p_{ik}^{n_{ik}},$$

   where n_{ik} is the frequency of class k at leaf i. Deviance is -2 times the log likelihood:

$$D = \sum_{i=1}^{v} D_i, \text{ where } D_i = -2 \sum_{k=1}^{c} n_{ik} \ln(p_{ik})$$

   and $0 \ln(0) = 0$. S-plus and R use deviance as a measure of leaf impurity.

(b) Gini index: An alternative measure of leaf impurity is the Gini index:

$$G = \sum_{i=1}^{v} G_i, \text{ where } G_i = \sum_{j \neq k}^{c} p_{ij} p_{ik} = 1 - \sum_{k=1}^{c} p_{ik}^2.$$

4. Splitting Rules

(a) S-plus and R allow binary splits only. If a split is made on a categorical variable having m levels, then there are $2^m - 1$ possible splits to consider. If a split is made on a numerical variable having m ordered values, then there are $m - 1$ possible splits of the form $y_i < t$ versus $y_i \geq t$.

(b) At each step the split that minimizes average impurity (over all leaves) is made. For example, if leaf i is split into leaf s and leaf t, then the deviance after the split is

$$D = \sum_{j=1}^{v} D_j - D_i + (D_s + D_t).$$

The decrease in deviance due to the split is

$$\begin{aligned} \text{Improvement} &= D_i - (D_s + D_t) \\ &= -2 \sum_{k=1}^{c} [n_{ik} \ln(p_{ik}) - n_{sk} \ln(p_{sk}) - n_{tk} \ln(p_{tk})] \\ &= -2 \sum_{k=1}^{c} \left[n_{sk} \ln\left(\frac{p_{ik}}{p_{sk}}\right) + n_{tk} \ln\left(\frac{p_{ik}}{p_{tk}}\right) \right]. \end{aligned}$$

Substituting MLEs $\widehat{p}_{ik} = n_{ik}/n_{i\cdot}$ yields

$$\text{Improvement} = -2 \sum_{k=1}^{c} \left[n_{sk} \ln\left(\frac{n_{ik} n_{s\cdot}}{n_{i\cdot} n_{sk}}\right) + n_{tk} \ln\left(\frac{n_{ik} n_{t\cdot}}{n_{i\cdot} n_{tk}}\right) \right].$$

Note that $-\log$ is a convex function. It is easy to show (using Jensen's inequality) that the improvement is non-negative.

(c) Splitting continues until the number of cases reaching each leaf is small (default in S-plus is 10) or if a leaf is homogeneous (leaf deviance is less than 1% of deviance at root node).

5. Missing Values: Drop a case down the tree as far as possible.

6. Pruning

(a) Strategy: Denote the deviance (or error rate) at leaf i by R_i and let $R = \sum_{i=1}^{v} R_i$. The number of leaves is taken to be the size of the tree. Breiman et al showed that the set of rooted subtrees of T that minimize a cost/complexity measure

$$R_\alpha = R - \alpha \times \text{ size}$$

is nested. That is, a nested set of subtrees is obtained by minimizing R_α for various values of α.

(b) Choice of α

i. AIC

ii. Cross-validation: S-plus will partition the data into 10 sets. A tree is constructed using nine of the sets and is evaluated using the hold out set. The process is repeated for each set and the results are averaged over the ten analyses.