Multivariate Statistical Analysis Notes

J S

2024.03.15

Preface

This is my personal study notes of Multivariate Statistical Analysis course, instructed by Wanlu Deng, at Tsinghua University, during the spring semester of 2024. The notes is formulated throughout the semester, so it's also a personal study profile. The course textbook is *Applied Multivariate Statistical Analysis*[a], although the course slides are used as the main resource. The last update is on 2024.7.14.

Initially the notes is prepared as the "cheat sheet" for the partly open-book final exam, but I find it helpful to type the formulae by hand and devise the interpretations on my own understanding. On the whole, the notes focus on applications and necessary intuition rather than rigorous theoretical proof. A few mathematical proofs of interest are attached in the appendix, and the rest are either straightforward or accessible on the Internet. Apart from my limited contributions, a large proportion of explanatory texts are copied from the course slides. Moreover, the notes is greatly inspired by v1ncent19's notes. Sincere thanks!

[a] R. A. Johnson & D. W. Wichern

Contents

1 Matrix Fundamentals

1.1 Linear Algebra

- Properties of trace:

$$\begin{aligned}
&\mathrm{tr}(c\mathbf{A}) = c\mathrm{tr}(\mathbf{A})\\
&\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})\\
&\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})\\
&\mathrm{tr}(\mathbf{B^{-1}AB}) = \mathrm{tr}(\mathbf{A})\\
&\mathrm{tr}(\mathbf{A_1A_2}\cdots\mathbf{A_n}) = \mathrm{tr}(\mathbf{A_2A_3}\cdots\mathbf{A_nA_1}) = \mathrm{tr}(\mathbf{A_nA_1A_2}\cdots\mathbf{A_{n-1}})\\
&\mathrm{tr}(\mathbf{AA}^T) = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2
\end{aligned}$$

- Maximization Lemma:

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{(\mathbf{x}^T\mathbf{d})^2}{\mathbf{x}^T\mathbf{Ax}} = \mathbf{d}^T\mathbf{A}^{-1}\mathbf{d}$$

$$\arg\max_{\mathbf{x}\neq\mathbf{0}} \frac{(\mathbf{x}^T\mathbf{d})^2}{\mathbf{x}^T\mathbf{Ax}} = c\mathbf{A}^{-1}\mathbf{d} \text{ (c is a constant)}$$

- Maximization on a unit sphere: If $\mathbf{A}$ is a positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and corresponding normalized eigenvectors $\mathbf{e}_1, \cdots, \mathbf{e}_p$,then

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{Ax}}{\mathbf{x}^T\mathbf{x}} = \lambda_1 \ (\mathbf{x} = \mathbf{e}_1)$$

$$\min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{Ax}}{\mathbf{x}^T\mathbf{x}} = \lambda_p \ (\mathbf{x} = \mathbf{e}_p)$$

$$\max_{\mathbf{x}\perp\mathbf{e}_l,\cdots,\mathbf{e}_k} \frac{\mathbf{x}^T\mathbf{Ax}}{\mathbf{x}^T\mathbf{x}} = \lambda_1 \ (\mathbf{x} = \mathbf{e}_{k+1})$$

- Cauchy Schwarz Inequality:

$$(\mathbf{x}^T\mathbf{y})^2 \leq (\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})$$

$$(\mathbf{x}^T\mathbf{y})^2 \leq (\mathbf{x}^T\mathbf{Ax})(\mathbf{y}^T\mathbf{A}^{-1}\mathbf{y}) \text{ (A is positive definite)}$$

- Spectral Decomposition:
  - Definition: If a $n \times n$ square matrix $\mathbf{A}$ is diagonalizable, the spectral decomposition of $\mathbf{A}$ is:

    $$\mathbf{A} = \mathbf{Q\Lambda Q}^{-1}$$

    * $\mathbf{Q}$ is invertible and $\mathbf{\Lambda}$ is diagonal.
  - Properties:
    * The diagonal entry of $\mathbf{\Lambda}$ is $\mathbf{A}$'s eigenvalues.
    * The i-th column of $\mathbf{Q}$ is the i-th eigenvector which belongs to the i-th diagonal entry.
    * Let $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$, $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n]$, then

    $$\mathbf{A} = \mathbf{Q\Lambda Q}^T = \lambda_1\mathbf{q}_1\mathbf{q}_1^T + \lambda_2\mathbf{q}_2\mathbf{q}_2^T + \cdots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T$$

    $$\mathbf{A}^k = \mathbf{Q\Lambda}^k\mathbf{Q}^T = \lambda_1^k\mathbf{q}_1\mathbf{q}_1^T + \lambda_2^k\mathbf{q}_2\mathbf{q}_2^T + \cdots + \lambda_n^k\mathbf{q}_n\mathbf{q}_n^T$$

1.2 Matrix Calculus

In matrix differentiation, we usually use denominator layout:

$$\left(\frac{\partial y}{\partial x}\right)_{ij} = \frac{\partial y_i}{\partial x_j}$$

Under denominator layout, we have

$$\begin{aligned}
\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= \mathbf{A}^T & \frac{\partial \mathbf{x}^T\mathbf{A}}{\partial \mathbf{x}} &= \mathbf{A} \\
\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} &= 2\mathbf{x} & \frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} \\
\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} &= |\mathbf{A}|\mathbf{A}^{-1} & \frac{\partial \mathrm{tr}(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} &= \mathbf{B}^T \\
\frac{\partial \mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})}{\partial \mathbf{A}} &= -\mathbf{A}^{-1}\mathbf{B}^T\mathbf{A}^{-1} & \frac{\partial \mathbf{X}^{-1}}{\partial t} &= -\mathbf{X}^{-1}\frac{\partial \mathbf{X}}{\partial t}\mathbf{X}^{-1}
\end{aligned}$$

1.3 Matrix Operations on Random Vector

Given a $n \times p$ data matrix $\mathbf{X}$, we have:

$$\begin{aligned}
&\bar{\mathbf{X}} = \frac{1}{n}\mathbf{X}^T\mathbf{1} && \mathbf{S} = \frac{1}{n-1}\mathbf{X}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X} \\
&\mathbf{D} = \mathbf{I} \odot \mathbf{S} = \mathrm{diag}(s_{11}, s_{22}, \cdots, s_{pp}) && \mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}} \\
&\mathbb{E}(\mathbf{A}\mathbf{X}\mathbf{B}^T) = \mathbf{A}\mathbb{E}(\mathbf{X})\mathbf{B}^T && \mathrm{Var}(\mathbf{A}\mathbf{X}) = \mathbf{A}\mathrm{Var}(\mathbf{X})\mathbf{A}^T \\
&\mathrm{Cov}(\mathbf{b}^T\mathbf{x}, \mathbf{c}^T\mathbf{x}) = \mathbf{b}^T\mathrm{Cov}(\mathbf{x}, \mathbf{x})\mathbf{c} = \mathbf{b}^T\mathbf{S}\mathbf{c}
\end{aligned}$$

2 Multivariate Normal Distribution

2.1 Definition

For a p-dimensional multivariate normal random vector $\underset{p\times 1}{\mathbf{x}}$ with mean $\underset{p\times 1}{\boldsymbol{\mu}}$ and covariance matrix $\underset{p\times p}{\boldsymbol{\Sigma}}$, the probability density function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

This is usually denoted as $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

2.2 Bivariate Normal Distribution

In the bivariate case, let σ_1, σ_2 be the variance and ρ be the covariance, then

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) + (\frac{y-\mu_2}{\sigma_2})^2\right]\right\}$$

$$Y|X \sim \mathcal{N}(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), (1-\rho^2)\sigma_2^2)$$

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$$

2.3 Maximum Likelihood Estimation

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T}{n}$$

Invariance property: Let $\hat{\theta}$ be the maximum likelihood estimate of θ, then $h(\hat{\theta})$ is the maximum likelihood estimate of $h(\theta)$.

The proof can be found in the appendix.

2.4 Linear Transformation

If $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

- For a $p \times 1$ vector $\mathbf{a}$: $\mathbf{a}^T\mathbf{X} \sim \mathcal{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$
- For a $q \times p$ matrix $\mathbf{A}$: $\mathbf{A}\mathbf{X} + \mathbf{a} \sim \mathcal{N}_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{a}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$
- For a $p \times p$ square matrix $\mathbf{A}$: $\mathbb{E}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$
- $(\mathbf{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi_p^2$

2.5 Marginal Distribution

All subsets of $\mathbf{X}$ are normally distributed. If we respectively partition $\mathbf{X}$, its mean $\boldsymbol{\mu}$, and its covariance matrix $\boldsymbol{\Sigma}$ to p and $n - p$ dimensions as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

then

- $\mathbf{X}_1$ is distributed as $\mathcal{N}_q(\boldsymbol{\mu}_1\boldsymbol{\Sigma}_{11})$
- $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent $\Leftrightarrow \boldsymbol{\Sigma}_{12} = \mathbf{0}$

2.6 Conditional Distribution

The distribution of $\mathbf{X_1}$ given $\mathbf{X_2} = \mathbf{x_2}$ is

$$\mathbf{X_1}|\mathbf{X_2} = \mathbf{x_2} \sim \mathcal{N}_p\left(\mu_\mathbf{1} + \boldsymbol{\Sigma}_\mathbf{12}\boldsymbol{\Sigma}_\mathbf{22}^{\mathbf{-1}}(\mathbf{x_2} - \mu_\mathbf{2}), \boldsymbol{\Sigma}_\mathbf{11} - \boldsymbol{\Sigma}_\mathbf{12}\boldsymbol{\Sigma}_\mathbf{22}^{\mathbf{-1}}\boldsymbol{\Sigma}_\mathbf{21}\right)$$

Note that

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & * \\ * & * \end{bmatrix}$$

2.7 Sampling Distribution

2.7.1 One-dimensional Case

Sample:

$$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Sampling Distribution:

$$\hat{\mu} = \bar{X} \sim \mathcal{N}(\mu, \frac{1}{n}\sigma^2)$$

$$\begin{aligned}\hat{\sigma^2} = S &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &\sim \frac{\sigma^2}{n-1}\chi^2_{n-1}\end{aligned}$$

2.7.2 Multi-dimensional Case

Sample:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Sampling Distribution:

$$\hat{\mu} = \bar{\mathbf{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} = \mathbf{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \\ &\sim \frac{1}{n-1}W_p(n-1, \boldsymbol{\Sigma})\end{aligned}$$

2.8 Wishart Distribution

If $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_m \sim^{i.i.d} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, then

$$\sum_{i=1}^{m}\mathbf{Z}_i\mathbf{Z}_i^T \sim W_p(m, \boldsymbol{\Sigma})$$

This is called $\sum_{i=1}^{m}\mathbf{Z}_i\mathbf{Z}_i^T$ follows a Wishart distribution with m degrees of freedom.

- Wishart distribution is the multidimensional case of chi-squared distribution.
- If $\mathbf{A}_1 \sim W_p(m_1, \boldsymbol{\Sigma})$, $\mathbf{A}_2 \sim W_p(m_2, \boldsymbol{\Sigma})$ and $\mathbf{A}_1$, $\mathbf{A}_2$ are independent, then $\mathbf{A}_1 + \mathbf{A}_2 \sim W_p(m_1 + m_2, \boldsymbol{\Sigma})$.
- If $\mathbf{A} \sim W_p(m, \boldsymbol{\Sigma})$, then $\mathbf{C}\mathbf{A}\mathbf{C}^T \sim W_p(m, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$.

2.9 Large-Sample Behavior

- Law of large numbers: Let $\mathbf{Y}_1, \mathbf{Y}_2, \cdot, \mathbf{Y}_n$ be independent observations from a population with mean $\mathbb{E}(\mathbf{Y}_i) = \boldsymbol{\mu}$, then
$$\bar{\mathbf{Y}} = \frac{1}{n}(\mathbf{Y}_1 + \mathbf{Y}_2 + \cdot + \mathbf{Y}_n)$$
converges in probability to $\boldsymbol{\mu}$ as n increases without bound.
- Central limit theorem: Let $\mathbf{X}_1, \mathbf{X}_2, \cdot, \mathbf{X}_n$ be independent observations from a population with mean $\boldsymbol{\mu}$ and finite covariance $\boldsymbol{\Sigma}$, then
$$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \text{ has an approximate } \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \text{ distribution}$$
for large sample sizes. Here n should also be large relative to p.
- $$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \text{ is approximately } \chi^2_p$$

2.10 Other Properties

- If $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and distributed as $\mathcal{N}_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathcal{N}_n(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ respectively, then
$$\begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix} \sim N_{m+n}\left(\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right) \tag{1}$$

- Even if $\mathbf{X}_1$ and $\mathbf{X}_2$ are Multivariate Normal Variables, and $\text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ may not be independent. Joint multivariate normal distribution is needed.

- Sufficient statistics: Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be a random sample from a multivariate normal population $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$. Then, $\bar{\mathbf{X}}$ and $\mathbf{S}$ are sufficient statistics. It means, for normal populations, all of the information about the model parameters is contained in $\bar{\mathbf{X}}$ and $\mathbf{S}$.

- The contour of $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$ are ellipsoids centered at $\boldsymbol{\mu}$ with axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$ $\Big((\lambda_i, \mathbf{e}_i)$ is an eigenvalue-eig

3 Hypothesis Test for Multivariate Normal Population

3.1 Test for One Population Mean

3.1.1 One-Dimensional Case

Sample:

$$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$$

Hypotheses:

$$H_0 : \mu = \mu_0 \leftrightarrow H_1 : \mu \neq \mu_0$$

Under H_0:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$$

$$\Longrightarrow \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \sim \mathcal{N}(0, 1)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \frac{1}{n-1}\sigma^2\chi^2_{n-1}$$

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

Test statistics:

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \sim t_{n-1}$$

$$T^2 \sim F_{1,n-1}$$

3.1.2 Multi-Dimensional Case

Sample:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Hypotheses:

$$H_0 : \boldsymbol{\mu} = \mu_{\mathbf{0}} \leftrightarrow H_1 : \boldsymbol{\mu} \neq \mu_{\mathbf{0}}$$

Under H_0:

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \sim \mathcal{N}_p(\mu_{\mathbf{0}}, \frac{1}{n}\boldsymbol{\Sigma})$$

$$\Longrightarrow \sqrt{n}(\bar{\mathbf{X}} - \mu_{\mathbf{0}}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$\sim \frac{1}{n-1}Wishart_p(n-1, \boldsymbol{\Sigma})$$

Test statistic: Hotelling's T^2:

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

$$F = \frac{n-p}{p}\frac{T^2}{n-1} \sim F_{p,n-p}$$

3.2 Likelihood Ratio Test

Sample:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \sim^{i.i.d} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Hypotheses:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \leftrightarrow H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Test statistic:

$$\boldsymbol{\Lambda} = \frac{\max_{H_0} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max_{H_0 \cup H_1} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}_0|}{|\hat{\boldsymbol{\Sigma}}|}\right)^{-\frac{n}{2}} = \left(1 + \frac{T^2}{n-1}\right)^{-\frac{n}{2}}$$

We have:

$$\frac{|\hat{\boldsymbol{\Sigma}}_0|}{|\hat{\boldsymbol{\Sigma}}|} = \frac{|\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_0)(\mathbf{X}_i - \boldsymbol{\mu}_0)^T|}{|\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})\mathbf{X}_i - \bar{\mathbf{X}})^T|}$$

$$T^2 = (n-1)\left(\frac{|\hat{\mathbf{\Sigma}}_0|}{|\hat{\mathbf{\Sigma}}|} - 1\right) = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \sim \frac{p(n-1)}{n-p} F_{p,n-p}$$

3.3 Test for Correlation Coefficient*

Sample:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}, \mathbf{\Sigma})$$

$$r = \frac{s_{12}}{\sqrt{s_{11}s_{22}}}$$

Hypotheses:

$$H_0 : \rho_{12} = 0 \Longleftrightarrow H_1 : \rho_{12} \neq 0$$

Under H_0:

$$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$$

3.4 Test for Single Anomalous Value*

Sample:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$$

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

Hypotheses:

$$H_0 : \mathbf{x}_i \text{ comes from the multivariate normal population} \Longleftrightarrow H_1 : \mathbf{x}_i \text{ is anomalous}$$

Under H_0:

$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^T \mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \text{ (Mahalanobis distance)}$$

$$F_i = \frac{(n-p-1)nD_i^2}{p(n-1)^2 - npD_i^2} \sim F_{p,n-p-1}$$

When there are multiple anomalous values, the diagnostics becomes intractable, because extreme values “hide” each other.

3.5 Simutaneous Confidence Region

3.5.1 Bonferroni Correction

Let $H_1, H_2, \cdots, H_m$ be a family of null hypotheses and let $p_1, p_2, \cdots, p_m$ be their corresponding p-values. The Bonferroni correction rejects the null hypothesis for each $p_i \leq \frac{\alpha}{m}$, thereby controlling the family-wise error rate(FWER) at $\leq \alpha$.

$$FWER \leq \sum_{i=1}^{m} P(p_i \leq \frac{\alpha}{m}) \leq \alpha$$

3.5.2 Confidence Interval of linear combination

- Sample:

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$$

$$z_i = \mathbf{a}^T\mathbf{X}_i \sim \mathcal{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\mathbf{\Sigma}\mathbf{a}) = \mathcal{N}(\phi, \mathbf{a}^T\mathbf{\Sigma}\mathbf{a})$$

- Hypothesis:

$$H_0 : \phi = \phi_0 \leftrightarrow H_1 : \phi \neq \phi_0$$

- Under H_0:

$$\bar{z} = \mathbf{a}^T\bar{\mathbf{X}} \sim \mathcal{N}(\phi_0, \frac{1}{n}\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}) \Longrightarrow \frac{\sqrt{n}(\bar{z}-\phi_0)}{\sqrt{\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}}} \sim \mathcal{N}(0,1)$$

$$s\{\hat{\phi}\} = \mathrm{se}(\bar{z}) = \frac{1}{\sqrt{n}}s_z = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2} = \sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}}$$

  where S is the sample variance of X and $s\{\cdot\}$ denotes standard deviation.

- Test statistic:

$$\left(\frac{\bar{z}-\phi_0}{s_z/\sqrt{n}}\right)^2 \sim F_{1,n-1}$$

- Confidence interval:
  - $100(1-\alpha)\%$ confidence interval (single):

$$I_\phi(z) = \hat{\phi} \pm t_{n-1}(\frac{\alpha}{2})\mathrm{se}(\hat{\phi}) = \mathbf{a}^T\bar{\mathbf{X}} \pm t_{n-1}(\frac{\alpha}{2})\sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}}$$

  - $100(1-\alpha)\%$ confidence interval with Bonferroni correction (m combinations in total):

$$I_{\phi_k}(z) = \hat{\phi_k} \pm t_{n-1}(\frac{\alpha}{2m})\hat{se}(\hat{\phi_k}) = \mathbf{a}^T\bar{\mathbf{X}} \pm t_{n-1}(\frac{\alpha}{2m})\sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}}$$

  - $100(1-\alpha)\%$ confidence interval (T^2):

$$I(z) = \mathbf{a}^T\bar{\mathbf{X}} \pm c\sqrt{\frac{\mathbf{a}^T\mathbf{S}\mathbf{a}}{n}}, \quad c^2 = T^2(\alpha) = \frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)$$

  The projection of T^2 confidence region on corresponding direction of $\mathbf{a}$ (the linear combination vector) is T^2 confidence interval.

3.6 Test for Two Population Mean

3.6.1 One-Dimensional Case (Under Equal Unknown Variance)

Sample:

$$X_{11}, X_{12}, \cdots, X_{1,n_1} \overset{i.i.d}{\sim} N(\mu_1, \sigma^2)$$

$$X_{21}, X_{22}, \cdots, X_{2,n_2} \overset{i.i.d}{\sim} N(\mu_2, \sigma^2)$$

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = \delta_0 \leftrightarrow H_1 : \mu_1 - \mu_2 \neq \delta_0$$

Under H_0:

$$\hat{\mu_1} - \hat{\mu_2} = \bar{X_1} - \bar{X_2} \sim \mathcal{N}(\delta_0, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$
$$\Longrightarrow \frac{\bar{X_1} - \bar{X_2} - \delta_0}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim \mathcal{N}(0,1)$$

$$(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 =$$
$$\sum_{i=1}^{n_1}(X_{1i} - \bar{X_1})^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X_2})^2 \sim \sigma^2 \chi^2_{n_1+n_2-2}$$
$$\Longrightarrow \frac{s_w^2}{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \sim \frac{1}{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \frac{\chi^2_{n_1+n_2-2}}{n_1 + n_2 - 2}$$

(Denote $\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ as s_w^2)

Test statistic:

$$T = \frac{\bar{X_1} + \bar{X_2} - \delta_0}{s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

3.6.2 Multi-Dimensional Case (Under Equal Unknown Variance)

Sample:

$$\mathbf{X}_{11}, \mathbf{X}_{12}, \cdots, \mathbf{X}_{1,n_1} \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$\mathbf{X}_{21}, \mathbf{X}_{22}, \cdots, \mathbf{X}_{2,n_2} \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

Hypotheses:

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \leftrightarrow H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0$$

Under H_0:

$$\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 \sim \mathcal{N}_p(\boldsymbol{\delta}_0, \frac{\boldsymbol{\Sigma}}{n_1} + \frac{\boldsymbol{\Sigma}}{n_2})$$
$$\Longrightarrow (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \delta_0)^T (\frac{\boldsymbol{\Sigma}}{n_1} + \frac{\boldsymbol{\Sigma}}{n_2})^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \delta_0) \sim \chi^2_p$$

$$(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 =$$
$$\sum_{i=1}^{n_1}(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)(\mathbf{X}_{1i} - \bar{\mathbf{X}}_1)^T + \sum_{i=1}^{n_2}(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)(\mathbf{X}_{2i} - \bar{\mathbf{X}}_2)^T$$
$$\sim \text{Wishart}_p(n_1 + n_2 - 2, \boldsymbol{\Sigma})$$

(Denote $\frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2)}{n_1 + n_2 - 2}$ as $\mathbf{S}_w$)

Test statistic:

$$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}_0)^T \left((\frac{1}{n_1} + \frac{1}{n_2}) \mathbf{S}_w \right)^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - \boldsymbol{\delta}_0)$$

$$F = \frac{n_1 + n_2 - p - 1}{p} \frac{T^2}{n_1 + n_2 - 2} \sim F_{p, n_1+n_2-p-1}$$

3.6.3 One-Dimensional Case (Paired Sample)

Sample:

$$X_{11}, X_{12}, \cdots, X_{1n} \overset{i.i.d}{\sim} N(\mu_1, \sigma_1^2)$$

$$X_{21}, X_{22}, \cdots, X_{2n} \overset{i.i.d}{\sim} N(\mu_2, \sigma_2^2)$$

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = \delta_0 \leftrightarrow H_1 : \mu_1 - \mu_2 \neq \delta_0$$

Under H_0:

$$\begin{aligned} z_i &= X_{1i} - X_{2i} \\ \Longrightarrow & z_1, z_2, \cdots, z_i \sim^{i.i.d} N(\delta_0, \sigma_1^2 + \sigma_2^2) \end{aligned}$$

This yields the one-population test regarding z.

$$\bar{z} \sim \mathcal{N}(\delta_0, \frac{\sigma_1^2 + \sigma_2^2}{n})$$

$$(n-1)s^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2 \sim (\sigma_1^2 + \sigma_2^2)\chi_{n-1}^2$$

Test statistic:

$$T = \frac{\sqrt{n}(\bar{z} - \delta_0)}{s} \sim t_{n-1}$$

3.6.4 Multi-Dimensional Case (Paired Sample)

Sample:

$$\mathbf{X}_{11}, \mathbf{X}_{12}, \cdots, \mathbf{X}_{1n} \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

$$\mathbf{X}_{21}, \mathbf{X}_{22}, \cdots, \mathbf{X}_{2n} \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

Hypotheses:

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \leftrightarrow H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0$$

Under H_0:

$$\begin{aligned} \mathbf{z}_i &= \mathbf{X}_{1i} - \mathbf{X}_{2i} \\ \Longrightarrow & \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_i \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\delta}_0, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \end{aligned}$$

This yields the one-population test regarding $\mathbf{z}$.

$$\bar{\mathbf{z}} \sim \mathcal{N}_p(\boldsymbol{\delta}_0, \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{n})$$

$$(n-1)\mathbf{S} = \sum_{i=1}^{n} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T \sim \text{Wishart}_p(n-1, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$$

Test statistic:

$$T^2 = n(\bar{\mathbf{z}} - \boldsymbol{\delta}_0)^T \mathbf{S}^{-1} (\bar{\mathbf{z}} - \boldsymbol{\delta}_0)$$

$$F = \frac{n-p}{p} \frac{T^2}{n-1} \sim F_{p,n-p}$$

3.6.5 General Case($n_1 \neq n_2, \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$)

See Behrens-Fisher Problem.

4 Principal Component Analysis

4.1 Population Principal Components

Consider a p-dimensional random variable $\mathbf{X} = (X_1, X_2, \cdots, X_p)^T$ with covariance matrix $\boldsymbol{\Sigma}$. The principal components are uncorrelated linear combinations whose variances are as large as possible.

$$\begin{aligned} \text{i-th principal component } &= \text{linear combination } \ \mathbf{a}_i^T\mathbf{X} \\ &\text{that maximizes } \ \text{Var}(\mathbf{a}_i^T\mathbf{X}) \\ &\text{subject to } \ \mathbf{a}_i^T\mathbf{a}_i = 1 \\ &\text{and } \ \text{Cov}(\mathbf{a}_i^T\mathbf{X}, \mathbf{a}_j^T\mathbf{X}) = 0 \ \text{ for } \ 1 \leq j < i \end{aligned}$$

Denote the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$ as $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \cdots, (\lambda_p, \mathbf{e}_p)$ $, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the spectral decomposition of covariance matrix $\boldsymbol{\Sigma}$ is

$$\underset{p\times p}{\boldsymbol{\Sigma}} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$$

where[a]

$$\underset{p\times p}{\mathbf{P}} = \begin{bmatrix} \underset{p\times 1}{\mathbf{e}_1}, \underset{p\times 1}{\mathbf{e}_2}, \cdots, \underset{p\times 1}{\mathbf{e}_p} \end{bmatrix}$$

$$\underset{p\times p}{\boldsymbol{\Lambda}} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p), \ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

The principal components of random variable $\mathbf{X}$ are

$$\begin{aligned} Y_1 &= \mathbf{e}_1^T\mathbf{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= \mathbf{e}_2^T\mathbf{X} = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{e}_p^T\mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

In matrix form, we denote principal components $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_p)^T$ as

$$\underset{p\times 1}{\mathbf{Y}} = \underset{p\times p}{\mathbf{P}^T} \underset{p\times 1}{\mathbf{X}}$$

4.1.1 Properties

- $\text{Var}(Y_i) = \mathbf{e}_i^T\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i, \quad \text{Cov}(Y_i, Y_j) = \mathbf{e}_i^T\boldsymbol{\Sigma}\mathbf{e}_j = 0$
- Let the diagonal entries of $\boldsymbol{\Sigma}$ be $\Sigma_{11}, \Sigma_{22}, \cdots, \Sigma_{pp}$, then

$$\sum_{i=1}^p \Sigma_{ii} = \sum_{i=1}^p \text{Var}(X_i) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \text{Var}(Y_i)$$

-

$$\rho_{Y_i,X_j} = \frac{e_{ij}\sqrt{\lambda_i}}{\sqrt{\Sigma_{jj}}} = \frac{e_{ij}\sqrt{\lambda_i}}{\sigma_{jj}}$$

$$\sum_{i=1}^p \rho^2_{Y_i,X_j} = 1$$

So the correlation coefficient is the proportion of the variance of X_k explained by the correlation with Y_i.

4.1.2 Principal Components Obtained from Standardized Variables

Principal Components can also be obtained from standardized variables $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)^T$ where $Z_i = (X_i - \mu_i)/\sqrt{\sigma_{ii}}$. In matrix notation,

$$\mathbf{Z} = \mathbf{V}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$$

where $\mathbf{V} = diag(\Sigma_{11}, \Sigma_{22}, \cdots, \Sigma_{pp})$ consists of the diagonal values of $\boldsymbol{\Sigma}$ (the covariance matrix of $\mathbf{X}$).

$\mathbf{Z}$ satisfies:

$$\mathbb{E}(\mathbf{Z}) = \mathbf{0}$$

$$\text{Cov}(\mathbf{Z}) = \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\Sigma}\mathbf{V}^{-\frac{1}{2}} = \boldsymbol{\rho}$$

The i-th principal component of the standardized variables $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)^T$ with $\text{Cov}(\mathbf{Z}) = \rho$ is given by

$$Y_i = \mathbf{e}_i^T\mathbf{Z} = \mathbf{e}_i^T\mathbf{V}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu})$$

where $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \cdots, (\lambda_p, \mathbf{e}_p)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are eigenvalue-eigenvector pairs of $\boldsymbol{\rho}$.

Moreover,

$$\sum_{i=1}^p \text{Var}(Y_i) = \sum_{i=1}^p \text{Var}(Z_i) = \sum_{i=1}^p \lambda_i = p$$

$$\rho_{Y_i,Z_j} = e_{ij}\sqrt{\lambda_i}$$

[a] $\mathbf{P}$ is an orthogonal matrix.

4.2 Sample Principal Component

Consider the p-dimensional data $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}$ drawn from a population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The data yield the sample mean vector $\bar{\mathbf{x}}$, sample covariance matrix $\mathbf{S}$, and sample correlation matrix $\mathbf{R}$.

Similarly, the i-th sample principal component is given by

$$\hat{y_i} = \hat{\mathbf{e_i}}^{\mathbf{T}}\mathbf{x} = \hat{e_{i1}}x_1 + \hat{e_{i2}}x_2 + \cdots + \hat{e_{ip}}x_p$$

where $\hat{\lambda_1} \geq \hat{\lambda_2} \geq \cdots \geq \hat{\lambda_p} \geq 0$ and $\mathbf{x}$ is an observation.

4.2.1 Properties

$$\text{Sample variance } \mathrm{Var}(\hat{y_i}) = \hat{\lambda}_i$$

$$\text{Total sample variance} = \sum_{i=1}^{p} S_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i$$

$$r_{\hat{y}_i, x_j} = \frac{\hat{e}_{ij}\sqrt{\hat{\lambda}_i}}{\sqrt{S_{jj}}} = \frac{\hat{e}_{ij}\sqrt{\hat{\lambda}_i}}{s_{jj}}$$

4.2.2 Sample Principle Components Obtained from Standardized Variables

For the sample, the standardization is accomplished by

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

The sample covariance matrix is

$$\mathbf{S}_z = \frac{1}{n}(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{Z})^T(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{Z}) = \mathbf{R}$$

$$\text{Sample variance } \mathrm{Var}(\hat{y_i}) = \hat{\lambda}_i$$

$$\text{Total (standardized) sample variance} = \sum_{i=1}^{p} \hat{\lambda}_i = p = \mathrm{tr}(\mathbf{R})$$

$$r_{\hat{y}_i, x_j} = \hat{e}_{ij}\sqrt{\hat{\lambda}_i}$$

4.3 Singular Value Decomposition

The principal components transformation can also be implemented by Singular Value Decomposition (SVD).

Denote $(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}$ as $\tilde{\mathbf{X}}$, then $\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$.

$$\begin{aligned} &\text{PCA: } \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{P}\boldsymbol{\Lambda}_0\mathbf{P}^T \\ &\text{SVD: } \tilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T \end{aligned}$$

All principal components $\underset{n\times p}{\mathbf{Y}} = (\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_p)^T$ are

$$\mathbf{Y} = \tilde{\mathbf{X}}\mathbf{P} = \tilde{\mathbf{X}}\mathbf{V} = \mathbf{U}\boldsymbol{\Lambda}$$

If $n \gg p$, PCA is faster; if $n \ll p$, SVD is faster. In terms of numerical stability, SVD is better than PCA.

4.4 Comments

- The general objectives of PCA are data reduction and interpretation.
- PCA is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.
- Principal components depend solely on the covariance matrix $\mathbf{\Sigma}$.
- If some $\hat{\lambda}_i$ is near 0, pay attention to linear dependency.
- Principal component is the projection of sample data $\mathbf{x}$ on eigen vector $\mathbf{e}$.
- Limitations:
  - Only uses information in the covariance matrix.
  - Only considers linear structure.
  - May be difficult to interpret linear combination of variables.
  - Sensitive to outliers.

5 Factor Analysis

5.1 Orthogonal Factor Model

Consider a p-dimensional random variable $\mathbf{X} = (X_1, X_2, \cdots, X_p)^T$ with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Factor analysis decomposes $\mathbf{X}$ into some internal factors:

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

In matrix form,

$$\underset{p\times 1}{\mathbf{X} - \boldsymbol{\mu}} = \underset{p\times m}{\mathbf{L}}\ \underset{m\times 1}{\mathbf{F}} + \underset{p\times 1}{\boldsymbol{\varepsilon}}$$

$\mathbf{L}$ is loading matrix; $\mathbf{F}$ is random internal factors; $\boldsymbol{\varepsilon}$ is random error.

Constraints:

$$\begin{aligned} &\mathbb{E}(\mathbf{F}) = \mathbf{0} \quad \mathrm{Cov}(\mathbf{F}) = \mathbb{E}(\mathbf{F}\mathbf{F}^T) = \underset{\mathbf{m\times m}}{\mathbf{I}} \\ &\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \underset{\mathbf{p\times p}}{\mathbf{\Psi}} = \mathrm{diag}(\psi_1, \psi_2, \cdots, \psi_p) \\ &\mathrm{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = \mathbb{E}(\boldsymbol{\varepsilon}\mathbf{F}^T) = \mathbf{0} \end{aligned}$$

Properties:

- Variance decomposition: The covariance matrix can be decomposed as:

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$$

  - Diagonal entries:

$$\mathrm{Var}(X_i) = \Sigma_{ii} = \sum_{k=1}^{m} l_{ik}^2 + \psi_i = h_i^2 + \psi_i$$

    where h_i^2 is communality, ψ_i is specific variance.

  - Nondiagonal entries:

$$\mathrm{Cov}(X_i, X_j) = \Sigma_{ij} = \sum_{k=1}^{m} l_{ik} l_{jk}$$

  - Correlation between X and F:

$$\mathrm{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$$

$$\mathrm{Cov}(X_i, F_j) = l_{ij}$$

- Nonidentifiability: $\mathbf{L}$ is not unique. For orthogonal matrix $\mathbf{T}$, $\mathbf{L}^* = \mathbf{LT}$ is also legal loading matrix, and $\mathbf{F}^* = \mathbf{T}^T\mathbf{F}$ is corresponding factors. Rotating $\mathbf{L}$ won't affect specific variances.

- Scale invariant: Consider a diagonal matrix $\mathbf{C} = \mathrm{diag}(c_1, c_2, \cdots, c_p)$.

$$\begin{aligned} \mathbf{X} &= \boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\varepsilon} \\ \mathbf{Y} &= \mathbf{CX} \\ &= \mathbf{C}\boldsymbol{\mu} + \mathbf{CLF} + \boldsymbol{\varepsilon}_c \\ &= \boldsymbol{\mu}_c + \mathbf{L}_c\mathbf{F} + \boldsymbol{\varepsilon}_c \end{aligned}$$

  So FA is unaffected by rescaling of the variables.

Limitations:

- Not all covariance matrix can be factored as $\mathbf{LL}^T + \boldsymbol{\Psi}$, where the number of factors $m \ll p$.

- Maybe the solution exists mathematically, but not statistically. e.g. correlation > 1 or variance < 0.

5.2 Principle Component Approach

- Objective: Find $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$, with $\hat{\mathbf{S}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\boldsymbol{\Psi}}$

- Solution: By spectral decomposition:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1 & \sqrt{\lambda_2}\mathbf{e}_2 & \cdots & \sqrt{\lambda_p}\mathbf{e}_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1^T \\ \sqrt{\lambda_2}\mathbf{e}_2^T \\ \vdots \\ \sqrt{\lambda_p}\mathbf{e}_p^T \end{bmatrix} = \mathbf{L}_0\mathbf{L}_0^T$$

  When the last $p - m$ eigenvalues are small, neglect the contribution of the corresponding eigenvalue-eigenvector pairs,

$$\boldsymbol{\Sigma} \approx \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1 & \sqrt{\lambda_2}\mathbf{e}_2 & \cdots & \sqrt{\lambda_m}\mathbf{e}_m \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1^T \\ \sqrt{\lambda_2}\mathbf{e}_2^T \\ \vdots \\ \sqrt{\lambda_m}\mathbf{e}_m^T \end{bmatrix} = \mathbf{LL}^T$$

  So

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 \quad \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 \quad \cdots \quad \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m \right]$$

  The specific variances may be taken as the diagonal elements of $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$.

$$\tilde{\boldsymbol{\Psi}} = \mathbf{I} \odot (\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T)$$

$$\psi_i = \Sigma_{ii} - \sum_{k=1}^{m} l_{ik}^2$$

- Variance interpretation: The proportion of total sample variance due to i-th factor is
$$\frac{\hat{\lambda}_i}{S_{11}+S_{22}+\cdots+S_{pp}}$$
For a factor ananysis of $\mathbf{R}$ (rather than $\mathbf{S}$), the proportion of total (standardized) sample variance due to i-th factor is
$$\frac{\hat{\lambda}_i}{p}$$

5.3 Maximum Likelihood Approach

- Assumption: Factor $\mathbf{F}$ and error ε are normally distributed:
$$\mathbf{F}\sim\mathcal{N}_m(\mathbf{0},\mathbf{I})\quad \varepsilon\sim\mathcal{N}_p(\mathbf{0},\mathbf{\Psi})\quad \mathbf{F}\perp\varepsilon$$
- Likelihood function:
$$L(\boldsymbol{\mu},\mathbf{\Sigma})=(2\pi)^{-\frac{np}{2}}|\mathbf{\Sigma}|^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}\left(n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})^T+\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})^T\right)\right]\right\}$$
The maximum likelihood estimate $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ must be obtained by numerical maximization.
- Constraints: For the uniqueness of $\mathbf{L}$, we usually assume
$$\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L}=\underset{\text{diagonal}}{\mathbf{D}}$$
Thus we get $\mathbf{L}$ and $\mathbf{\Psi}$ by
$$\mathbf{\Sigma}=\mathbf{L}\mathbf{L}^T+\mathbf{\Psi}$$
- Comparison: The cumulative proportion of the total sample variance explained by the factors is larger for principal component factoring than for maximum likelihood factoring, due to the variance maximizing design of PCA.

5.4 Standardization

Denote $\mathbf{Z}=\mathbf{V}^{-\frac{1}{2}}(\mathbf{X}-\boldsymbol{\mu})$, where $\mathbf{V}=\mathrm{diag}(\mathbf{\Sigma})=\mathrm{diag}(\Sigma_{11},\Sigma_{22},\cdots,\Sigma_{pp})$.

Normal version	Standardized version
$\mathbf{X}-\boldsymbol{\mu}=\mathbf{L}\mathbf{F}+\boldsymbol{\varepsilon}$	$\mathbf{Z}=\mathbf{L}_z\mathbf{F}+\boldsymbol{\varepsilon}_z$
$\mathbf{V}^{-\frac{1}{2}}(\mathbf{X}-\boldsymbol{\mu})=\mathbf{V}^{-\frac{1}{2}}\mathbf{L}\mathbf{F}+\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}$	$\mathbf{V}^{-\frac{1}{2}}(X-\boldsymbol{\mu})=\mathbf{L}_z\mathbf{F}+\boldsymbol{\varepsilon}_z$

So
$$\mathbf{L}_z=\mathbf{V}^{-\frac{1}{2}}\mathbf{L},\ \ \mathbf{\Psi}_z=\mathbf{V}^{-\frac{1}{2}}\mathbf{\Psi}\mathbf{V}^{-\frac{1}{2}}$$
For estimation,
$$\hat{\mathbf{L}}_z=\hat{\mathbf{V}}^{-\frac{1}{2}}\hat{\mathbf{L}},\ \ \hat{\mathbf{\Psi}}_z=\hat{\mathbf{V}}^{-\frac{1}{2}}\hat{\mathbf{\Psi}}\hat{\mathbf{V}}^{-\frac{1}{2}}$$

5.5 Factor Rotation

- Intuition: Due to the nonidentifiability property, loading matrix $\mathbf{L}$ is not unique. Rotating $\mathbf{L}$ won't affect specific variances and communalities.

  Since the original loadings may not be readily interpretable, it is usual practice to rotate them until a simpler structure is attained. Ideally, we want a pattern of loadings that each variable loads highly on a single factor and has small to moderate loadings on the remaining factors.

- Varimax criterion: Define $\widetilde{l}^*_{ij} = \frac{\hat{l}^*_{ij}}{\hat{h}_i}$ to be the rotated coefficients scaled by communalities.
  Kaiser proposes an analytical measure of simple structure:

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}(\widetilde{l}^*_{ij})^4 - \frac{\left(\sum_{i=1}^{p}\widetilde{l}^{*2}_{ij}\right)^2}{p}\right]$$

  Maximizing V corresponds to “spreading out” the squares of loadings on each factor as much as possible.

- Advanced topics: The method above is orthogonal rotation, which means the factors are uncorrelated (under multivariate normal assumption, independent). But there is also “oblique rotation” where factors are not necessarily uncorrelated. An oblique rotations corresponds to a nonrigid rotation of the coordinate system such that the rotated axes pass through the clusters, but are not perpendicular to one another.

5.6 Factor Scores

- Definition: Factor scores are estimates of values for the unobserved random factor vectors $\mathbf{F_j}$, $j = 1, 2, \cdots, n$. That is, factor scores

$$\hat{\mathbf{f}}_j = \text{estimate of the values } \mathbf{f}_j \text{ attained by } \mathbf{F}_j \text{ (j-th case)}$$

- Solution: Suppose the mean vector $\boldsymbol{\mu}$, the factor loadings $\mathbf{L}$, and the specific variance $\boldsymbol{\Psi}$ are known for the factor model

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$$

  The form is similar to linear regression, except that the variances ε are unequal. With weighted least squares approach, the sum of squared errors can be written as

$$\sum_{i=1}^{p}\frac{\varepsilon_i^2}{\psi_i} = \boldsymbol{\varepsilon}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} = (\mathbf{x} - \boldsymbol{\mu} - \mathbf{LF})^T\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu} - \mathbf{LF})$$

  The solution is

$$\hat{\mathbf{f}} = (\mathbf{L}^T\boldsymbol{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}^T\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

  If we take the estimates $\hat{\mathbf{L}}$, $\hat{\boldsymbol{\Psi}}$ and $\boldsymbol{\mu} = \bar{\mathbf{x}}$ as the true values, we obtain the factor scores for the j-th case as

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}^T\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}^T\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

5.6.1 Weighted Least Squares Method

- MLE approach: When $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ are determined by maximum likelihood approach, which is $\mathbf{L}^T\boldsymbol{\Psi}^{-1}\mathbf{L} = \mathbf{D}$,

$$\hat{\mathbf{f}}_j = \mathbf{D}^{-1}\hat{\mathbf{L}}^T\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$$

  For the standardized version, or, if we take factor analysis based on the correlation matrix $\mathbf{R}$,

$$\hat{\mathbf{f}}_j = \mathbf{D}_z^{-1}\hat{\mathbf{L}}_z^T\hat{\boldsymbol{\Psi}}_{\mathbf{z}}^{-1}\mathbf{z}_j$$

- PC approach: If the factor scores are estimated by the PC approach, it is customary to generate factor scores with ordinary (unweighted) least squares procedure, which means we view specific variances ψ_i as nearly equal. The factor scores are

$$\hat{\mathbf{f}}_j = (\widetilde{\mathbf{L}}^T\widetilde{\mathbf{L}})^{-1}\widetilde{\mathbf{L}}^T(\mathbf{x}_j - \bar{\mathbf{x}})$$

  For the standardized version,

$$\hat{\mathbf{f}}_j = (\widetilde{\mathbf{L}}_z^T\widetilde{\mathbf{L}}_z)^{-1}\widetilde{\mathbf{L}}_{\mathbf{z}}^T\mathbf{z}_j$$

Recall that

$$\widetilde{\mathbf{L}} = \left[\sqrt{\hat\lambda_1}\hat{\mathbf{e}}_1 \quad \sqrt{\hat\lambda_2}\hat{\mathbf{e}}_2 \quad \cdots \quad \sqrt{\hat\lambda_m}\hat{\mathbf{e}}_m\right]$$

For the factor scores,

$$\frac{1}{n}\sum_{j=1}^{n}\hat{\mathbf{f}}_j = \mathbf{0} \text{ (sample mean)}$$

$$\frac{1}{n-1}\sum_{j=1}^{n}\hat{\mathbf{f}}_j\hat{\mathbf{f}}_j^T = \mathbf{I}$$

5.6.2 Regression method

We treat the loading matrix $\mathbf{L}$ and specific variance matrix $\mathbf{\Psi}$ as known. When the common factors $\mathbf{F}$ and specific factors $\boldsymbol{\varepsilon}$ are jointly normally distributed, we have

$$\begin{bmatrix} \underset{p\times 1}{\mathbf{X}-\boldsymbol{\mu}} \\ \underset{m\times 1}{\mathbf{F}} \end{bmatrix} \sim N_{p+m}\left(\begin{bmatrix} \underset{p\times 1}{\mathbf{0}} \\ \underset{m\times 1}{\mathbf{0}} \end{bmatrix}, \begin{bmatrix} \underset{p\times p}{\mathbf{L}\mathbf{L}^T+\mathbf{\Psi}} & \underset{p\times m}{\mathbf{L}} \\ \underset{m\times p}{\mathbf{L}^T} & \underset{m\times m}{\mathbf{I}} \end{bmatrix}\right)$$

The conditional distribution $\mathbf{F}|\mathbf{x}$ is

$$\mathbf{F}|\mathbf{x} \sim N_m\left(\mathbf{L}^T(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi})^{-1}(\mathbf{x}-\boldsymbol{\mu}), \mathbf{I}-\mathbf{L}^T(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi})^{-1}\mathbf{L}\right)$$

The j-th factor score vector is given by

$$\begin{aligned}\hat{\mathbf{f}}_j &= \hat{\mathbf{L}}^T(\hat{\mathbf{L}}\hat{\mathbf{L}}^T+\hat{\mathbf{\Psi}})^{-1}(\mathbf{x}_j-\bar{\mathbf{x}}) \\ &= (\mathbf{I}+\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}(\mathbf{x}_j-\bar{\mathbf{x}})\end{aligned}$$

If we denote the factor score obtained by least squares method and regression as $\hat{\mathbf{f}}_j^{LS}$ and $\hat{\mathbf{f}}_j^{R}$ separately, we have

$$\hat{\mathbf{f}}_j^{LS} = \left[\mathbf{I}+(\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}})^{-1}\right]\hat{\mathbf{f}}_j^{R}$$

For MLE estimates, $(\hat{\mathbf{L}}^T\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}})^{-1} = \mathbf{D}^{-1}$. If the elements of the diagonal matrix $\mathbf{D}^{-1}$ is near zero, the regression method gives nearly the same answer as least squares method.

So the factor scores obtained by regression is:

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}^T\mathbf{S}^{-1}(\mathbf{x}_j-\bar{\mathbf{x}})$$

For the standardized version,

$$\hat{\mathbf{f}}_j = \hat{\mathbf{L}}_{\mathbf{z}}^T\mathbf{R}^{-1}\mathbf{z}_j$$

5.6.3 Comparison between two methods

- Unbiasedness:

$$\begin{aligned}\mathbf{f}^{LS} &= (\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{\Psi}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ \mathbf{f}^{R} &= \mathbf{L}^T(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi})^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ \mathbb{E}(\mathbf{f}^{LS}|\mathbf{F}) &= (\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L}\mathbf{F} = \mathbf{F} \text{ (unbiased)} \\ \mathbb{E}(\mathbf{f}^{R}|\mathbf{F}) &= \mathbf{L}^T(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi})^{-1}\mathbf{L}\mathbf{F} = \left(\mathbf{I}+(\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\right)^{-1}\mathbf{F} \text{ (biased)}\end{aligned}$$

- Mean prediction error:

$$\begin{aligned}\mathbf{f}^{LS} &= (\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{\Psi}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{F}+(\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^T\mathbf{L}^T\mathbf{\Psi}\boldsymbol{\varepsilon} \\ \mathbf{f}^{R} &= \mathbf{L}^T(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi})^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{L}^T(\mathbf{L}\mathbf{L}^T+\mathbf{\Psi})^{-1}(\mathbf{L}\mathbf{F}+\boldsymbol{\varepsilon}) \\ \mathbb{E}\left((\mathbf{f}^{LS}-\mathbf{F})(\mathbf{f}^{LS}-\mathbf{F})^T)|\mathbf{F}\right) &= (\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1} \text{ (larger)} \\ \mathbb{E}\left((\hat{\mathbf{f}}^{R}-\mathbf{F})(\hat{\mathbf{f}}^{R}-\mathbf{F})^T)|\mathbf{F}\right) &= (\mathbf{I}+\mathbf{L}^T\mathbf{\Psi}^{-1}\mathbf{L})^{-1} \text{ (smaller)}\end{aligned}$$

In terms of unbiaseness, least squares method is better; in terms of mean prediction error, regression method is better.

5.7 Large Sample Test for the Number of Common Factors*

- Sample: n random observations of dimension p:

$$\underset{p\times1}{\mathbf{x}_1}, \underset{p\times1}{\mathbf{x}_2}, \cdots, \underset{p\times1}{\mathbf{x}_n}$$

- Hypotheses:

$$H_0 : \mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T + \mathbf{\Psi} \Longleftrightarrow H_1 : \mathbf{\Sigma} \text{ is any other positive definite matrix}$$

- Test statistic:

$$\chi^2 = (n - 1 - \frac{2p + 4m + 5}{6}) \ln \frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}}|}{|\mathbf{S}_n|}$$

  - $\mathbf{S}_n = \hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$
  - $\nu = \dfrac{(p-m)^2 - (p+m)}{2}$
  - m is the number of internal factors.

- Distribution:

$$\chi^2 \overset{H_0}{\sim} \chi^2_\nu$$

- Decision rule: If $\chi^2 > \chi^2(\alpha)$, reject H_0; Otherwise, accept H_0.

5.8 Comments

- The essential purpose of factor analysis is to describe the covariance structure among many variables with a few unobservable or latent variables called factors.
  - Reduction: reduce high dimension data to a few variables.
  - Interpretation: explain the covariance of observed variables with latent factors.
- Note that the original covariance matrix has $\frac{p(p+1)}{2}$ parameters, while FA model has only $p(m+1)$ parameters. (pm from $\mathbf{L}$ and p from $\mathbf{\Psi}$)
- When the number of factors(m) changes, the estimated loading for a given factor doesn't change.
- Comparison between FA and PCA:
  - FA assumes that the covariance is $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T + \mathbf{\Phi}$, while PCA doesn't make assumptions of the covariance structure.
  - When the variance of unique factor is 0, principal factor analysis is equivalent to PCA. When the variance of unique factor is small, PCA is similar to FA; when large, the difference is significant.
  - PCA emphasizes the transformation from original variables to principal components, while FA emphasizes the transformation from underlying factors to original variables.

6 Canonical Correlation Analysis

Key idea: Canonical correlation analysis (CCA) is concerned with explaining the covariance structure of two sets of variables through a few linear combinations of these variables.

The general objectives are

- Identify and quantify the associations between two sets of variables.
- Data/Dimension reduction while retaining original information in covariance as much as possible.

6.1 Canonical Variates and Canonical Correlations

- Sample: Consider two random variables denoted by $\underset{p\times 1}{\mathbf{X}^{(1)}}$ and $\underset{q\times 1}{\mathbf{X}^{(2)}}$ $(p \leq q)$.

  Assume

  $$\underset{(p+q)\times 1}{\mathbf{X}} = \begin{bmatrix} \underset{p\times 1}{\mathbf{X}^{(1)}} \\ \underset{q\times 1}{\mathbf{X}^{(2)}} \end{bmatrix}$$

  has mean vector

  $$\underset{(p+q)\times 1}{\boldsymbol{\mu}} = \mathbb{E}(\mathbf{X}) = \begin{bmatrix} \mathbb{E}(\mathbf{X}^{(1)}) \\ \mathbb{E}(\mathbf{X}^{(2))} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

  and covariance matrix

  $$\underset{(p+q)\times(p+q)}{\boldsymbol{\Sigma}} = \mathbb{E}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T = \begin{bmatrix} \underset{p\times p}{\boldsymbol{\Sigma}_{11}} & \underset{p\times q}{\boldsymbol{\Sigma}_{12}} \\ \underset{q\times p}{\boldsymbol{\Sigma}_{21}} & \underset{q\times q}{\boldsymbol{\Sigma}_{22}} \end{bmatrix}$$

- Objective: In CCA, we summarize the association between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ with linear combinations. The target is to find a pair of coefficient vectors $\underset{p\times 1}{\mathbf{a}}$ and $\underset{p\times 1}{\mathbf{b}}$, such that $\mathrm{Corr}(U, V)$ is as large as possible, where

  $$U = \mathbf{a}^T\mathbf{X}^{(1)}, \ \ V = \mathbf{b}^T\mathbf{X}^{(2)}$$

  Thus we obtain

  $$\mathrm{Var}(U) = \mathbf{a}^T\boldsymbol{\Sigma}_{11}a, \ \ \mathrm{Var}(V) = \mathbf{b}^T\boldsymbol{\Sigma}_{22}\mathbf{b}$$

  $$\mathrm{Cov}(U,V) = \mathbf{a}^T\boldsymbol{\Sigma}_{12}\mathbf{b}, \ \ \mathrm{Corr}(U,V) = \frac{\mathbf{a}^T\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}^T\boldsymbol{\Sigma}_{22}b}}$$

  The optimization objective is

  $$\max_{\mathbf{a},\mathbf{b}\neq 0} \frac{\mathbf{a}^T\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}^T\boldsymbol{\Sigma}_{22}\mathbf{b}}}$$

  Or

  $$\max_{|\tilde{\mathbf{a}}|=|\tilde{\mathbf{b}}|=1} \tilde{\mathbf{a}}^T\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\tilde{\mathbf{b}}$$

- Concepts: Generally, denote the i-th pair of canonical variates as

  $$U_i = \mathbf{a}_i^T\mathbf{X}^{(1)}, \ \ V_i = \mathbf{b}_i^T\mathbf{X}^{(2)}$$

  U_i, V_i are attained by maximizing $\mathrm{Corr}(U_i, V_i)$ or $\frac{\mathbf{a}_i^T\boldsymbol{\Sigma}_{12}\mathbf{b}_i}{\sqrt{\mathbf{a}_i^T\boldsymbol{\Sigma}_{11}\mathbf{a}_i}\sqrt{\mathbf{b}_i^T\boldsymbol{\Sigma}_{22}\mathbf{b}_i}}$ where U_i, V_i are uncorrelated with the previous $i-1$ pairs.

  The i-th canonical correlation is

  $$\rho_i^* = \mathrm{Corr}(U_i, V_i) = \max_{\mathbf{a}_i,\mathbf{b}_i} \frac{\mathbf{a}_i^T\boldsymbol{\Sigma}_{12}\mathbf{b}_\mathbf{i}}{\sqrt{\mathbf{a}_i^T\boldsymbol{\Sigma}_{11}\mathbf{a}_i}\sqrt{\mathbf{b}_i^T\boldsymbol{\Sigma}_{22}\mathbf{b}_i}}$$

  $$\begin{aligned} \text{subject to } \ \mathrm{Var}(U_i) &= \mathbf{a}_i^T\boldsymbol{\Sigma}_{11}\mathbf{a}_i = 1 \\ \mathrm{Var}(V_i) &= \mathbf{b}_i^T\boldsymbol{\Sigma}_{22}b_i = 1 \\ \mathrm{Cov}(U_i, U_k) &= \mathbf{a}_i^T\boldsymbol{\Sigma}_{11}\mathbf{a}_k = 0, \ \ \forall \ k < i \\ \mathrm{Cov}(V_i, V_k) &= \mathbf{b}_i^T\boldsymbol{\Sigma}_{22}\mathbf{b}_k = 0, \ \ \forall \ k < i \end{aligned}$$

- Solution: The i-th canonical variate pair is

  $$U_i = \mathbf{e}_i^T\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{X}^{(1)}, \ \ V_i = \mathbf{f}_i^T\boldsymbol{\Sigma}_{22}^{-\frac{1}{2}}\mathbf{X}^{(2)}$$

- $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \rho_p^{*2}$ are the eigenvalues of $\mathbf{\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}}$, and $\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_p}$ are the corresponding eigenvectors.
- $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \rho_p^{*2}$ are the eigenvalues of $\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}}$, and $\mathbf{f_1}, \mathbf{f_2}, \cdots, \mathbf{f_p}$ are the corresponding eigenvectors. Each $\mathbf{f_i}$ is proportional to $\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i}$

The canonical variates have the properties:

$$\mathrm{Var}(U_i) = \mathrm{Var}(V_i) = 1$$

$$\mathrm{Cov}(U_i, U_j) = \mathrm{Corr}(U_i, U_j) = 0 \ (i \neq j)$$
$$\mathrm{Cov}(V_i, V_j) = \mathrm{Corr}(V_i, V_j) = 0 \ (i \neq j)$$
$$\mathrm{Cov}(U_i, V_j) = \mathrm{Corr}(U_i, V_j) = 0 \ (i \neq j)$$

for $i, j \in \{1, 2, \cdots, p\}$.

- Matrix form: Reformulate all pairs of canonical variates as

$$\underset{p\times p}{\mathbf{A}} = [\mathbf{a_1}, \mathbf{a_2}, \cdots, \mathbf{a_p}]^T, \quad \underset{q\times q}{\mathbf{B}} = [\mathbf{b_1}, \mathbf{b_2}, \cdots, \mathbf{b_q}]^T$$

Then all the canonical variates can be expressed by

$$\underset{p\times 1}{\mathbf{U}} = \mathbf{AX^{(1)}}, \quad \underset{q\times 1}{\mathbf{V}} = \mathbf{BX^{(2)}}$$

Thus we have the correlations between all the canonical variates and the original variables $\mathbf{X^{(1)}}$ and $\mathbf{X^{(2)}}$:

$$\mathrm{Corr}(\mathbf{U}, \mathbf{X^{(1)}}) = \mathbf{A\Sigma_{11}V_{11}^{-\frac{1}{2}}} \quad \mathrm{Corr}(\mathbf{U}, \mathbf{X^{(2)}}) = \mathbf{A\Sigma_{12}V_{22}^{-\frac{1}{2}}}$$
$$\mathrm{Corr}(\mathbf{V}, \mathbf{X^{(1)}}) = \mathbf{B\Sigma_{21}V_{11}^{-\frac{1}{2}}} \quad \mathrm{Corr}(\mathbf{V}, \mathbf{X^{(2)}}) = \mathbf{B\Sigma_{22}V_{22}^{-\frac{1}{2}}}$$

where

$$\mathbf{V_{11}} = \mathrm{diag}(\mathbf{\Sigma_{11}}), \quad \mathbf{V_{22}} = \mathrm{diag}(\mathbf{\Sigma_{22}})$$

6.2 CCA with Standardized Variables

Canonical variates can be obtained for the standardized variables:

$$Z_k^{(j)} = \frac{X_k^j - \mu_k^{(j)}}{\sqrt{\Sigma_{kk}^{(j)}}}, \ j = 1, 2; \ k = 1, 2, \cdots, p$$

$$\mathbf{Z}^{(j)} = \mathbf{V}_{jj}^{-\frac{1}{2}}(\mathbf{X}^{(j)} - \boldsymbol{\mu}^{(j)})$$

Replace the covariance matrix with correlation matrix:

$$\underset{(p+q)\times(p+q)}{\rho} = \begin{bmatrix} \underset{p\times p}{\boldsymbol{\rho}_{11}} & \underset{p\times q}{\boldsymbol{\rho}_{12}} \\ \underset{q\times p}{\boldsymbol{\rho}_{21}} & \underset{q\times q}{\boldsymbol{\rho}_{22}} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{11}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{22}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{11}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{22}^{-\frac{1}{2}} \end{bmatrix} \quad (\mathbf{V}_{jj} = \mathrm{diag}(\mathbf{\Sigma}_{jj}))$$

Similarly, the i-th canonical variate pair is

$$U_i = \mathbf{e}_i^T \boldsymbol{\rho}_{11}^{-\frac{1}{2}} \mathbf{Z}^{(1)}, \quad V_i = \mathbf{f}_i^T \boldsymbol{\rho}_{22}^{-\frac{1}{2}} \mathbf{Z}^{(2)}$$

- $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \rho_p^{*2}$ are the eigenvalues of $\boldsymbol{\rho}_{11}^{-\frac{1}{2}}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-\frac{1}{2}}$, and $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$ are the corresponding eigenvectors.
- $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \rho_p^{*2}$ are the eigenvalues of $\boldsymbol{\rho}_{22}^{-\frac{1}{2}}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-\frac{1}{2}}$, and $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_p$ are the corresponding eigenvectors.

6.3 Large Sample Inference

If $\mathbf{\Sigma}_{12} = \mathbf{0}$, then all canonical correlations are 0, and it's meaningless to do CCA.
To test whether $\mathbf{\Sigma}_{12} = \mathbf{0}$, follow the procedure:

- Hypothesis:
$$H_0 : \rho_1^* = \rho_2^* = \cdots = \rho_p^* = 0 \ \ (\Leftrightarrow \mathbf{\Sigma_{12}} = 0)$$
- Test statistic:
$$\chi^2 = -2\ln\Lambda = n\ln\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|} = -n\ln\prod_{i=1}^{p}(1-\hat{\rho}_i^{*2})$$
- Distribution:
$$\chi^2 \overset{H_0}{\to} \chi^2_{pq}$$

7 Discriminant Analysis

Objective: Discriminant analysis, also known as pattern recognition/classification, is concerned with separating distinct sets of objects. The goal is to determine a rule to assign a new object (with unknown category) to one of the pre-specified categories, based on known observations.

7.1 Classification for Two Populations

Suppose a p-dimensional random variable $\underset{p\times 1}{\mathbf{X}}$.

- Notations:
  - Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density function from the two populations π_1 and π_2, respectively.
  - Let Ω be the sample space (the collection of all possible observations of $\mathbf{x}$).
  - Let R_1 be the region in which we classify objects as π_1 and R_2 be the remaining region in which we classify objects as π_2.
  - Let p_1 and p_2 be the prior probabilities of π_1 and π_2, where $p_1 + p_2 = 1$.
- Cost: The cost of misclassification can be defined by a cost matrix:

True label	Classify as	
	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

- ECM: Expected cost of misclassification.
$$\begin{aligned} ECM &= c(2|1)P(\mathbf{X}\in R_2, \mathbf{X}\in\pi_1) + c(1|2)P(\mathbf{X}\in R_2, \mathbf{X}\in\pi_1) \\ &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \end{aligned}$$
where $P(i|j) = P(\mathbf{X}\in R_i|\mathbf{X}\in\pi_j)$.
The target is to find R_1 that minimizes ECM.
- ECM region: By minimizing ECM, we get
$$R_1 : \left\{\mathbf{X}\Big|\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\}$$
$$R_2 : \left\{\mathbf{X}\Big|\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right\}$$
Special cases:

- Equal prior probabilities:

$$R_1 : \left\{ \mathbf{X} \Big| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \right\} \quad R_2 : \left\{ \mathbf{X} \Big| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)} \right\}$$

- Equal misclassification costs:

$$R_1 : \left\{ \mathbf{X} \Big| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \right\} \quad R_2 : \left\{ \mathbf{X} \Big| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \right\}$$

- Both equal prior probabilities and equal misclassification costs:

$$R_1 : \left\{ \mathbf{X} \Big| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \right\} \quad R_2 : \left\{ \mathbf{X} \Big| \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1 \right\}$$

- TPM: Total Probability of Misclassification.

$$TPM = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$$

The posterior probability is

$$\begin{aligned} P(\pi_1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|\pi_1)P(\pi_1)}{P(\mathbf{x}_0|\pi_1)P(\pi_1) + P(\mathbf{x}_0|\pi_2)P(\pi_2)} &= \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)} \\ P(\pi_2|\mathbf{x}_0) = 1 - P(\pi_1|\mathbf{x}_0) &= \frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)} \end{aligned}$$

The problem is equivalent to minimizing ECM when the cost of misclassification is equal $\Big(c(2|1) = c(1|2)\Big)$.

$$\arg\min \text{TPM} = \arg \min_{c(2|1)=c(1|2)} \text{ECM}$$

7.2 Classification with Two Multivariate Normal Populations

Assume $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities:

$$\pi_1 : N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad \pi_2 : N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}^{\frac{1}{2}}|} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right] \quad (i = 1, 2)$$

Based on the central limit theorem, multivariate normal assumption is useful in statistical pratice.

7.2.1 Equal Variance

By minimizing ECM, we get the regions

$$\begin{aligned} R_1 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right] \\ R_2 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < \ln\left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right] \end{aligned}$$

In most practical situations, the population quantities $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ are unknown.

Suppose we have n_1 and n_2 observations from π_1 and π_2, respectively.
The estimate of expectation is the sample mean:

$$\hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{x}}_1 \quad \hat{\boldsymbol{\mu}}_2 = \bar{\mathbf{x}}_2$$

The estimate of variance is the pooled version:

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

Replace $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ by $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_p$ to get the regions R_1 and R_2.

7.2.2 Unequal Variance

By minimizing ECM, we get the regions

$$R_1 : -\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

$$R_2 : -\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1})\mathbf{x} - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

where

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$$

The classification region are defined by quadratic functions of $\mathbf{x}$.

When the population quantities $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ are unknown, replace them by $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1, \mathbf{S}_2$ to get R_1 and R_2.

7.3 Fisher LDA

7.3.1 Two populations

- Idea: Transform the multivariate observation $\mathbf{x}$ to univariate observations y such that the y's derived from population π_1 and π_2 are separated as much as possible.
- Assuption: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. (No normality assumption)
- Sample: A set of p-dimensional random variables, categorized as π_1 and π_2, with n_1 and n_2 observations from each category, sample mean $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$ and sample variance $\mathbf{S}_1, \mathbf{S}_2$.
- Objective:

$$\max_{\mathbf{a}} \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

  - s_y^2 is the pooled estimate: $s_y^2 = \dfrac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$
  - $\mathbf{a}$ is the transformation vector: $y = \mathbf{a}^T\mathbf{x}$.
- Solution: Denote the pooled sample variance as $\mathbf{S}_p = \frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2}{n_1+n_2-2}$, then

$$a \propto \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$\max_{\mathbf{a}} \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

7.3.2 Several populations

- Assuption: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$.
- Sample: A set of p-dimensional random variables, categorized as $\pi_1, \pi_2, \cdots, \pi_g$, with $n_1, n_2, \cdots, n_g$ observations from each category, sample mean $\bar{\mathbf{x}}_1, \cdots, \bar{\mathbf{x}}_g$ and sample variance $\mathbf{S}_1, \cdots, \mathbf{S}_g$.
- Objective:

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$$

  - $\mathbf{B} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})^\mathbf{T}$ is the between group variance.
  - $\mathbf{W} = \sum_{i=1}^{g}(n_i - 1)\mathbf{S}_\mathbf{i}$ is the within group variance.

- Solution: Let $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_s$ denote the $s \leq \min(g-1, p)$ nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, and $\hat{\mathrm{e}}_1, \cdots, \hat{\mathbf{e}}_s$ denote the correponding eigenvectors (scaled to $\hat{\mathbf{e}}_i^T \mathbf{S}_p \hat{\mathbf{e}}_i = 1$).

  The $\mathbf{a}$ to maximize $\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$ is given by
  $$\hat{\mathbf{a}}_1 = \hat{\mathbf{e}}_1$$
  The linear combination $\hat{\mathbf{a}}_1^T\mathbf{x}$ is called the sample first discriminant. Similarly, $\hat{\mathbf{a}}_i^T\mathbf{x}$ is called the sample i-th discriminant, with $\hat{\mathbf{a}}_i = \hat{\mathbf{e}}_i$.

- Classification rule: Denote $Y_k = \mathbf{a}_k^T\mathbf{X}$ as the k-th discriminant ($k < s$). Each observation $\mathbf{X}$ is transformed as $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_s]^T$. Also, denote $\boldsymbol{\mu}_{iY}$ be the mean vector of $\mathbf{Y}$ for samples from the i-th group.

  To classify, assign $\mathbf{y}$ to the i-th group if $\mathbf{y}$ is closest to the mean of the i-th group in Euclidean distance.

7.4 Classification for Several Populations

In g population case ($g \geq 3$), ECM (Expected cost of misclassification) is given by

$$\mathrm{ECM} = \sum_{i=1}^{g} \sum_{j=1; j\neq i}^{g} p_i P(i|i) c(j|i)$$

When c's are all equal, the regions that minimize ECM are

$$R_k : \left\{ \mathbf{X} \middle| p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}), \forall i \neq k \right\}$$

Equivalent result is the posterior distribution:

$$P(\mathbf{X} \in \pi_k | \mathbf{X} = \mathbf{x}_0) = \frac{p_k f_k(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0) + \cdots + p_g f_g(\mathbf{x}_0)}$$

7.5 Evaluation of Classification

Here we take two population case as example.

- OER: Optimum error rate.
  $$\mathrm{OER} = p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$$
  $\hat{R}_1$ and $\hat{R}_2$ represent the regions determined by probability distribution functions of each group.

- AER: Actual error rate.
  $$\mathrm{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x}) d\mathbf{x}$$
  $\hat{R}_1$ and $\hat{R}_2$ represent the regions determined by observed samples.

- APER: Apparent error rate, the proportion of items in the training set that are misclassified.

  APER tends to underestimate AER, because the data used to build the classification function are also used to evaluate it (overfitting). To mitigate the problem, we can use cross validation.

8 Cluster Analysis

8.1 Agglomorative Clustering Algorithm

```
Agglomorative Clustering Algorithm
Require: $n$ samples: $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}^p$
Require: $\mathrm{dist}(\cdot,\cdot)$: Binary distance function to measure the distance between two clusters.
Initialization: Let each data point be a cluster; Compute the distance matrix.
  while more than one cluster remains
    do
    1.Merge the two closest clusters.
    2.Update the distance matrix.
  end while
```

We can use dendrograms to record/visualize the clustering procedure.

8.1.1 Measurements of Distance

- Single linkage: The minimum distance of two clusters.

$$\min_{a\in A, b\in B} \mathrm{dist}(a,b)$$

  - Can handle irregularly shaped regions fairly naturally.
  - Sensitive to noise and outliers in the form of “chain”.

- Complete linkage: The maximum distance of two clusters.

$$\max_{a\in A, b\in B} \mathrm{dist(a,b)}$$

  - Less sensitive to noise and outliers than single linkage.
  - Regions are generally compact, but may violate “closeness”. That is, points may much closer to some points in neighbouringcluster than its own cluster.
  - This manifests itself as breaking large clusters.
  - Clusters are biased to be globular.

- Average linkage: The distance of the centroids of two clusters.

$$\min_{a\in A, b\in B} \mathrm{dist}(\bar{a},\bar{b})$$

  Shares globular clusters of complete, less sensitive than single.

- Ward’s linkage: Within cluster Sum of Squared Error (SSE):

$$SSE = \sum_{i=1}^{K} \sum_{x\in C_i} \mathrm{dist}^2(m_i, x)$$

  - Similar to average if dissimilarity between points is distance squared. So it shares many properties of average linkage.
  - A hierarchical analogue of K-means.
  - Sometimes used to initialize K-means.

8.1.2 Strengths and Limitations

Strengths:

- Do not have to assume the number of clusters:
  - Any desired number of clusters can be obtained by "cutting" the dendrogram at the proper level.
- Intuitive to display relationship among objects:
  - E.g. phylogenetic tree, which biologists use to illustrate the evolutionary relationship among species.

Limitations:

- Computational issues: High computational complexity.
  - $O(n^2)$ space. A proximity matrix is used.
  - $O(n^3)$ time. In many cases there are n steps, and at each step a matrix of size n^2 must be updated and/or searched.
- Statistical issues:
  - Once a decision is made to combine two clusters, it cannot be undone.
  - No objective function is directly minimized.
  - Different schemes have problems with one or more of the following:
    - Sensitivity to noise and outliers.
    - Difficulty to handle different sized clusters and convex shapes.
    - Breaking large clusters.

8.2 K-means Clustering Algorithm

```
K-means Clustering Algorithm
Require: $n$ samples: $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}^p$
Require: $\mathrm{dist}(\cdot,\cdot)$: Binary distance function to measure the distance between two points.
Require: $k$: The number of clusters.
Initialization: Choose the centroids of the $k$ clusters randomly/by certain rule.
  while not converge
    do
    1.For all points, calculate the distance from each centroid.
    2.Assign each point to the nearest cluster.
    3.Update the centroids of each cluster.
  end while
```

8.2.1 Choice of k

- Elbow method: calculate within cluster SSE for each k, use the turning point in the SSE-k curve.
- Empirical method: $k \approx \sqrt{n/2}$.
- Cross validation: Use cross validation to evaluate clustering quality for each k, and pick the best one.

8.2.2 Extensions of K-Means

- K-Modes: Replacing means of clusters with modes.
  - Improvement: can handle categorical data.
- K-Medoids: Use medoids (the most centrally located object in a cluster) instead of centroids.
  - Improvement: overcome the sensitivity to outliers.

8.2.3 Strengths and Limitations

Strength: efficient. Computational complexity is $O(tkn)$, where n is objects, k is clusters, and t is iterations. Normally, $k, t \ll n$.

Limitations:

- K-means is sensitive to outliers.
- K-means has problems whenclusters are of:
  - Different sizes.
  - Different densities.
  - Non-convex shape.

8.3 Gaussian Mixture Model with Expectation Maximization Algorithm

8.3.1 GMM Model

Assume there are k clusters, n observations, and each observation is p-dimensional. The Gaussian Mixture Model(GMM) assumes $\mathbf{X}$ is generated from a mixed distribution of k multivariate normal distributions.

- Assume π is a k-dimensional prior probability vector, where $\sum_{i=1}^{k} \pi_i = 1$. $\mathbf{x}$ has probability π_ℓ to be generated from corresponding normal $N(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$
- Assume there is an underlying label $Y \sim \text{Multinomial}(\ell, \boldsymbol{\pi})$, and $X|Y = \ell \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$. In real practice, we only observe $\mathbf{X}$ but don't know the underlying true labels Y.
- Assume $\gamma_{i\ell}$ is the posterior probability that the i-th object belongs to the ℓ-th group. All posterior probabilities consturct an $n \times p$ matrix $\boldsymbol{\Gamma}$.
- Goal: Estimate $\boldsymbol{\pi}, (\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell), 1 \leq \ell \leq k$
- Likelihood function: Assume $\underset{n\times p}{\mathbf{X}} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix}^T$, then

$$\begin{aligned} f(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \sum_{\ell=1}^{k} \pi_\ell f_l(\mathbf{x}_i | \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell) \\ &= \text{const} \prod_{i=1}^{n} \sum_{\ell=1}^{k} \pi_\ell |\boldsymbol{\Sigma}_\ell|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_\ell)^T \boldsymbol{\Sigma}_\ell^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_\ell) \right] \end{aligned}$$

  It's difficult to solve the parameters directly. We use expectation maximization algorithm to solve the problem.

8.3.2 EM Algorithm

```
Expectation Maximization Algorithm
Require: $n$ samples: $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}^p$
Require: $k$: The number of clusters.
Initialization: Initialize $\boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ randomly/by certain rule.
  while not converge
   do
   1.Estimate $\boldsymbol{\Gamma}$ (Expectation):
     $\hat{\gamma}_{i\ell}^{(t+1)} = P(Y_i = l|\mathbf{X}_i = \mathbf{x}_i; \hat{\boldsymbol{\pi}}^{(t)}, \hat{\boldsymbol{\mu}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)})$
     $= \frac{\hat{\pi}_\ell^{(t)} \phi(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_\ell^{(t)}, \hat{\boldsymbol{\Sigma}}_\ell^{(t)})}{\sum_{j=1}^k \hat{\pi}_j^{(t)} \phi(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_\ell^{(t)}, \hat{\boldsymbol{\Sigma}}_j^{(t)})}$
   2.Estimate $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ (Maximization).
     $\hat{\boldsymbol{\mu}}_\ell^{(t+1)} = \frac{1}{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)}} \sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)} \mathbf{x}_i$
     $\hat{\boldsymbol{\Sigma}}_\ell^{(t+1)} = \frac{1}{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)}} \sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_\ell)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_\ell)^T$
     $\hat{\pi}_\ell^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)}$
  end while
```

8.3.3 Comparison of LDA and GMM

LDA	GMM
Hard, based on observed Y	Soft, based on guessed Y
$\hat{\pi}_\ell = \frac{\sum_{i=1}^n I\left(Y_i = \ell\right)}{n}$	$\hat{\pi}_\ell^{(t+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)}}{n}$
$\hat{\mu}_\ell = \frac{\sum_{i=1}^n I\left(Y_i = \ell\right) \cdot x_i}{\sum_{i=1}^n I\left(Y_i = \ell\right)}$	$\hat{\mu}_\ell^{(t+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)} \cdot x_i}{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)}}$
$\hat{\Sigma}_\ell = \frac{\sum_{i=1}^n I\left(Y_i = \ell\right) \cdot \left(x_i - \hat{\mu}_\ell\right)\left(x_i - \hat{\mu}_\ell\right)^T}{\sum_{i=1}^n I\left(Y_i = \ell\right)}$	$\hat{\Sigma}_\ell^{(t+1)} = \frac{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)} \cdot \left(x_i - \hat{\mu}_\ell^{(t+1)}\right)\left(x_i - \hat{\mu}_\ell^{(t+1)}\right)^T}{\sum_{i=1}^n \hat{\gamma}_{i\ell}^{(t+1)}}$

8.3.4 Comments

- The quantities $\boldsymbol{\Gamma}$ are not real parameters, they are “estimates” of the random labels Y which were unobserved.
- If we had observed Y then the rows of $\boldsymbol{\Gamma}$ would be all zero except one entry being 1.
- The EM simply replaces the unobserved Y with a guess.

9 Multidimensional Scaling*

Reference: Peking University - Multivariate Statistical Analysis.

9.1 Classical Multidimensional Scaling

9.1.1 Background

Suppose we have n points in p-dimensional space. By some measure of distance, we have the distance matrix $\underset{n\times n}{\mathbf{D}}$ $(\mathbf{D} = \mathbf{D}^T, d_{ij} \geq 0, d_{ii} = 0, \forall i, j \in 1, 2, \cdots, n, i \neq j)$. We want to reduce the dimension of the data points with the distances preserved.

For instance, we want to reduce the dimension from p to k (usually $k = 1, 2, 3$), and get a sample matrix:

$$\underset{n\times k}{\mathbf{X}} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Denote the new distance matrix as $\hat{\mathbf{D}} = \left(\hat{d}_{ij}\right)_{n\times n}$, we want to make $\hat{\mathbf{D}}$ similar to $\mathbf{D}$.

9.1.2 Solution

```
Multidimensional Scaling Algorithm
Require: $n$ samples: $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \in \mathbb{R}_p$
Require: $k$: Target dimension.
Require: $\mathbf{D}$: Distance matrix for samples in $\mathbb{R}_p$.
1. Calculate $\mathbf{B}$:
   $\mathbf{B} = -\frac{1}{2}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$
2. Eigen decomposition:
   $\mathbf{B} = \mathbf{P\Lambda P^T}$
      * $\mathbf{P} = [\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_n}]$
      * $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$
      * $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$
3. Dimension Reduction:
   $\underset{k\times k}{\tilde{\mathbf{\Lambda}}} = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k) \quad (\lambda_1 \geq \lambda_2 \geq \cdots \lambda_k \geq 0)$
   $\underset{n\times k}{\tilde{\mathbf{P}}} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]$
   $\underset{n\times k}{\mathbf{X}} = \tilde{\mathbf{P}}\tilde{\mathbf{\Lambda}}^{\frac{1}{2}}$
```

9.1.3 Proof

Assume we have already known the transformed sample matrix $\mathbf{X}$. Let $\mathbf{B} = \mathbf{X}\mathbf{X}^T$, then $b_{ij} = \mathbf{x}_i^T\mathbf{x}_j$, and we get the association between b_{ij} and d_{ij}.

$$d_{ij}^2 = \sum_{l=1}^{k}(x_{il} - x_{jl})^2 = \sum_{l=1}^{k} x_{il}^2 + \sum_{l=1}^{k} x_{jl}^2 - 2\sum_{l=1}^{k} x_{il}x_{jl} = b_{ii} + b_{jj} - 2b_{ij}$$

Since position shift doesn't influence the distance matrix, we set each $\mathbf{x}_i$ to be centered at 0 on each coordinate.

$$\text{Constraint: } \sum_{i=1}^{n} x_{il} = 0 \ \ (l = 1, 2, \cdots, k)$$

This yields

$$\sum_{i=1}^{n} b_{ij} = 0, \quad \sum_{j=1}^{n} b_{ij} = 0$$

Thus we get

$$\sum_{i=1}^{n} d_{ij}^2 = T + nb_{jj} \quad \sum_{j=1}^{n} d_{ij}^2 = T + nb_{ii} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nT \quad T = \sum_{i=1}^{n} b_i i = \mathrm{tr}(B)$$

So we get the iterative expression of b_{ij}:

$$b_{ij} = (b_{ii} + b_{ij} - d_{ij}^2)/2$$

And we can further derive

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2)$$

where $d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$, $d_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2$, $d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n} d_{ij}^2$.

In neat matrix notation, we have

$$\mathbf{B} = -\frac{1}{2}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$$

By eigen decomposition,

$$\mathbf{B} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$$

Choose the first k eigenvalues and corresponding eigenvectors, we get $\underset{n\times k}{\tilde{\mathbf{P}}}$, $\underset{k\times k}{\tilde{\mathbf{\Lambda}}}$, and transformed $\mathbf{X}$:

$$\underset{n\times k}{\mathbf{X}} = \tilde{\mathbf{P}}\tilde{\mathbf{\Lambda}}^{\frac{1}{2}}$$

10 Appendix

Collection of some (relatively) unimportant but interesting topics.

10.1 Proof of Maximum Likelihood Estimation of Multivariate Normal Population

For multivariate normal distribution with dimension p,

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

On given observations $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, the log likelihood is:

$$\ln L(\boldsymbol{\mu}, \mathbf{\Sigma}) = -\frac{np}{2}\ln 2\pi - \frac{n}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

To maximize $\ln L$, we take partial deriatives on $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ and set them to 0:

$$\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\mu}, \mathbf{\Sigma})}{\partial \boldsymbol{\mu}} &= -\frac{1}{2}\sum_{i=1}^{n}\left[\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + (\mathbf{\Sigma}^{-1})^T(\mathbf{x}_i - \boldsymbol{\mu})\right] ^{\mathsf{b}} \\
&= -\mathbf{\Sigma}^{-1} \cdot \sum_{i=1}^{n}[(\mathbf{x}_i - \boldsymbol{\mu})] \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} &= -\frac{n}{2} \cdot \frac{1}{|\boldsymbol{\Sigma}|} \cdot |\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \sum_{i=1}^{n} \operatorname{tr}\left[(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right] \text{[a]} \\
&= -\frac{n}{2} \cdot \frac{1}{|\boldsymbol{\Sigma}|} \cdot |\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \sum_{i=1}^{n} \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\right] \\
&= -\frac{n}{2} \cdot \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \text{[b]} \\
&= -\frac{n}{2} \cdot \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \cdot \boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T\right] \boldsymbol{\Sigma}^{-1} \\
&= 0
\end{aligned}$$

So

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \mathbf{x}_i}{n}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T}{n}$$

[b] $\dfrac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = Ax + A^T x$

[a] $\dfrac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}|\mathbf{A}^{-1}$

[b] $\dfrac{\partial \operatorname{tr}(\mathbf{A}^{-1}\mathbf{B})}{\partial \mathbf{B}} = -\mathbf{A}^{-1}\mathbf{B}^T\mathbf{A}^{-1}$