COURSE CONTENT

1 Matrix manipulation
2 Multivariate data
3 Multivariate Normal Distribution
4 Mean Vector Inference
5 Principal Component Analysis
6 Factor Analysis
7 Discriminant Analysis
8 Cluster Analysis

Today: main points and additional revision problems. See course notes for full assessable content and the relevant textbook sections for further examples and problems.

1 MATRIX MANIPULATION

- Vector multiplication
- Addition
- Length $L_x = |x| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$

- Matrix transpose $\mathbf{X}'$ or $\mathbf{X}^{\mathrm{T}}$
- Multiplication
- Symmetric $\mathbf{X}' = \mathbf{X}$
- Identity matrix
- Inverse $\mathbf{X}^{-1} = \mathbf{AX} = \mathbf{XA} = \mathbf{I}$
- Trace
- Determinant $\begin{vmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{c} & \mathbf{d} \end{vmatrix} = ad - bc$ $\begin{vmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \\ \mathbf{d} & \mathbf{e} & \mathbf{f} \\ \mathbf{g} & \mathbf{h} & \mathbf{i} \end{vmatrix} = a(ei - fh) - b(di - fg) + c(dh - eg)$

- Eigenvalues & Eigenvectors (inc. normalised)

$$|\mathbf{A} - \boldsymbol{\lambda}\mathbf{I}| = \mathbf{0} \qquad \mathbf{A}e = \boldsymbol{\lambda}e \qquad e_i^T e_i = 1$$

1 RANDOM VECTORS

- Population mean vector μ
- Population covariance matrix Σ
- Population correlation matrix ρ

- Generalised variance $|\mathbf{\Sigma}|$

- Properties of Random Vectors

- Linear combinations

2 MULTIVARIATE DATA

- Random sample
- Sample mean vector $\overline{\mathbf{X}}_{\mathbf{n}}$
- Sample variance
  - Biased $$\mathbf{S}^*_{\mathbf{ii}} = \frac{1}{n}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_{in}\right)^2$$
  - unbiased $$\mathbf{S}_{\mathbf{ii}} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_{in}\right)^2$$
- Sample covariance
  - Biased
  - Unbiased

- Generalised variance
- Sample correlation

$$R_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}}$$

3 MULTIVARIATE NORMAL DISTRIBUTION

Notation $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \Sigma)$

Probability density function

$$f_p(\boldsymbol{y} : \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \boldsymbol{e}^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}-\boldsymbol{\mu})} \text{ for } \boldsymbol{y} \in \Re^p$$

Bivariate normal distribution

Properties of Multivariate Normal Distribution

Sampling Distribution

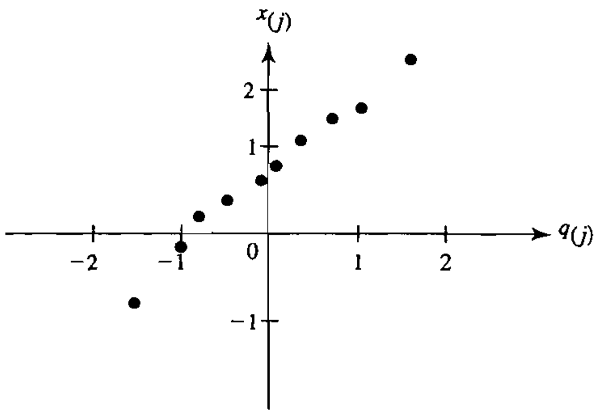
Hotelling T^2 Distribution $T^2 = n(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})^T \mathcal{S}_n^{-1} (\overline{\boldsymbol{X}}_n - \boldsymbol{\mu})$

Mahalanobis Distance $D_i^2 = (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^T \mathcal{S}_n^{-1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)$

3 MULTIVARIATE NORMAL DISTRIBUTION

Assessing Normality

Normal Probability Plot (Q-Q plot)



Large sample

central limit theorem

4 INFERENCES ABOUT THE MEAN VECTOR

Hypothesis testing

T^2

$$\frac{n-p}{p(n-1)}T^2 > F_{p,n-p}(\alpha) \qquad T^2 > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$$

Confidence interval

$$n\left(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}\right)^T \mathcal{S}_n^{-1}\left(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}\right) \leq \frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)$$

Simultaneous CI

$$\left(\overline{X}_{in} - \sqrt{\frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{S_{ii}}{n}}, \quad \overline{X}_{in} + \sqrt{\frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{S_{ii}}{n}}\right)$$

4 INFERENCES ABOUT THE MEAN VECTOR

Bonferroni Intervals

$$\left(\overline{X}_{in} - t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{S_{ii}}{n}}, \quad \overline{X}_{in} + t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{S_{ii}}{n}}\right)$$

Large sample and the use of chi-square χ_p^2

T^2 $$T^2 = n\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}_0\right)^T \mathcal{S}_n^{-1}\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}_0\right) \quad \approx \quad \chi_p^2.$$

CI $$n\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right)^T \mathcal{S}_n^{-1}\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right) \leq \chi_p^2(\alpha)$$

Simultaneous $$\left(\overline{X}_{in} - \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{S_{ii}}{n}}, \quad \overline{X}_{in} + \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{S_{ii}}{n}}\right)$$

5 PRINCIPAL COMPONENT ANALYSIS

Normalised linear combinations of correlated random variables

data reduction

transformation into independent variables

First component, maximum variance

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

Suppose $(\lambda_1, \boldsymbol{e}_1)$, $(\lambda_2, \boldsymbol{e}_2), \ldots,$ $(\lambda_p, \boldsymbol{e}_p)$ are the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, then, j^{th} pc is defined by

$$Y_j = \boldsymbol{e}_j^T \boldsymbol{X} \qquad\qquad \text{for } j = 1, 2, \ldots, p.$$

Then $\mathbf{Var}(Y_j) = \boldsymbol{e}_j^T \boldsymbol{\Sigma} \boldsymbol{e}_j = \lambda_j$ and $\mathbf{Cov}(Y_i, Y_j) = \boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_j = 0.$

5 PRINCIPAL COMPONENT ANALYSIS

Total variance

$$T = \sum_{i=1}^{n} \mathbf{Var}(X_i) = \sum_{i=1}^{n} \sigma_{ii} = tr(\mathbf{\Sigma}) = \lambda_1 + \lambda_2 + \ldots + \lambda_p,$$

Proportion

$$p_j = \lambda_j / T$$

Correlation coefficient Y_j and X_k

$$\rho_{Y_j, X_k} = \frac{e_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}} \qquad j, k = 1, 2, \ldots, p$$

$$\text{where } Y_j = e_{j1}X_1 + e_{j2}X_2 + \ldots + e_{jk}X_k + \ldots + e_{jp}X_p$$

Using correlation matrix

$$\begin{aligned} U_j &= \boldsymbol{\omega}_j^T \boldsymbol{Z} = \omega_{j1}Z_1 + \omega_{j2}Z_2 + \ldots + \omega_{jp}Z_p \\ &= \sum_i \omega_{ji} Z_i = \sum_i \omega_{ji} \left(\frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \right) \qquad \text{for } j = 1, 2, \ldots, p. \end{aligned}$$

5 PRINCIPAL COMPONENT ANALYSIS

Sample principal components

Inference in principal components

$$\left(\frac{\widehat{\lambda}_i}{1+a}, \frac{\widehat{\lambda}_i}{1-a}\right) \qquad\qquad a = z_{\alpha/2}\sqrt{\frac{2}{n}}$$

Smaller PC significance

$$H_0 : \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_{r+s} \quad \text{against}$$

$$H_1 : \lambda_{r+1}, \lambda_{r+2}, \cdots, \lambda_{r+s} \text{ are not all equal.}$$

Application to Linear Regression

6 FACTOR ANALYSIS

Orthogonal factor model

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon}$$

Assumptions

$$\mathbf{E}(\boldsymbol{F}) = \mathbf{0} \text{ and } \mathbf{Cov}(\boldsymbol{F}) = \boldsymbol{I}_m$$

$$\mathbf{E}(\varepsilon) = \mathbf{0} \text{ and } \mathbf{Cov}(\varepsilon) = \boldsymbol{\Psi}_p = \operatorname{diag}(\psi_1, \psi_2, \cdots, \psi_p)$$

$$\boldsymbol{F} \text{ and } \varepsilon \text{ are independent}$$

$$\Sigma = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}_p \text{ and}$$

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

6 FACTOR ANALYSIS

Principal Factor Method $\widehat{\boldsymbol{L}} = (\sqrt{\widehat{\lambda}_1}\widehat{\boldsymbol{e}}_1, \sqrt{\widehat{\lambda}_2}\widehat{\boldsymbol{e}}_2, \cdots, \sqrt{\widehat{\lambda}_m}\widehat{\boldsymbol{e}}_m)_{p\times m}$

Variance criterion $\kappa = 100\frac{\sum_{i=1}^{m}\widehat{\lambda}_i}{\sum_{i=1}^{p}\widehat{\lambda}_i}$

Eigenvalue criterion

Maximum Likelihood

Assumptions

(a) $\boldsymbol{F}$ is normal with $\mathbf{E}(\boldsymbol{F}) = \mathbf{0}$ and $\mathbf{Cov}(\boldsymbol{F}) = \boldsymbol{I}_m$. That is, $F_i \sim N(0,1)$ and mutually independent.

(b) ε is normal with $\mathbf{E}(\varepsilon) = \mathbf{0}$ and $\mathbf{Cov}(\varepsilon) = \boldsymbol{\Psi}_p$= diag $(\psi_1, \psi_2, \cdots, \psi_p)$. This implies that $\varepsilon_i \sim N(0, \psi_i)$ and mutually independent.

(c) $\boldsymbol{F}$ and ε are independent

$$\mathcal{L}(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}p}|\Sigma|^{\frac{n}{2}}} \exp\left\{\frac{1}{2}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu})^T\Sigma^{-1}\boldsymbol{x}_j - \boldsymbol{\mu})\right\}$$

unique $\boldsymbol{\Gamma} = \boldsymbol{L}^T\boldsymbol{\Psi}_p^{-1}\boldsymbol{L}$

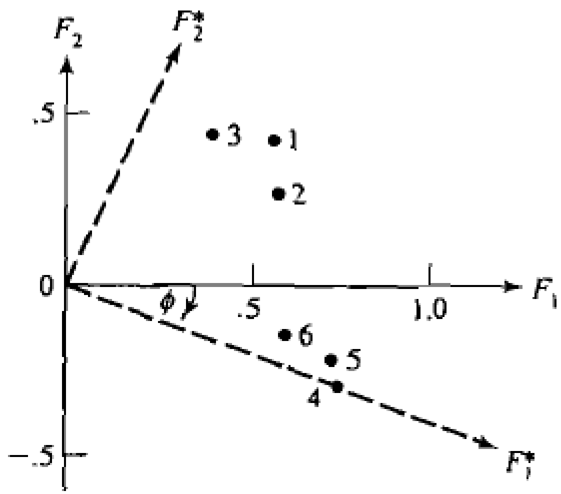
6 FACTOR ANALYSIS

Test for common factors

$$U = k \min\left(\mathcal{F}(\boldsymbol{L}, \Psi_p)\right)$$

$$m < \frac{1}{2}\left(2p + 1 - \sqrt{8p + 1}\right)$$

Factor Rotation



7 DISCRIMINANT ANALYSIS

Classification of Two Populations

Cost of misclassification

		Classified as:	
		Π_1	Π_2
True	Π_1	0	$c(2\|1)$
population	Π_2	$c(1\|2)$	0

$$\text{ECM} = c(2|1)p(2|1)p_1 + c(1|2)p(1|2)p_2$$

Allocate $\boldsymbol{x}_0$ to Π_1 if $\dfrac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \geq \dfrac{c(1|2)}{c(2|1)} \dfrac{p_2}{p_1}$

otherwise allocate $\boldsymbol{x}_0$ to Π_2.

When Normal and $\Sigma_1 = \Sigma_2 = \Sigma$ population and sample

When Normal and $\Sigma_1 \neq \Sigma_2$ population and sample

Evaluation APER and AER $\text{APER} = \dfrac{n_{1m} + n_{2m}}{n_1 + n_2}$ $\hat{\text{E}}(\text{AER}) = \dfrac{n_{1m}^{(H)} + n_{2m}^{(H)}}{n_1 + n_2}$

8 CLUSTER ANALYSIS

Minimum distance

Euclidean

$$D_{i,j} = ||\boldsymbol{x}_i - \boldsymbol{x}_j|| = \left\{ (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right\}^{1/2} = \left\{ \sum_{s=1}^{p} (x_{i,s} - x_{j,s})^2 \right\}^{1/2}$$

L_1

$$L_{i,j} = |\boldsymbol{x}_i - \boldsymbol{x}_j| = \sum_{s=1}^{p} |x_{i,s} - x_{j,s}|$$

K-means method $\widehat{\boldsymbol{\mu}}_r = \frac{1}{n_r} \sum_{\boldsymbol{x}_t \in C_r} \boldsymbol{x}_t$

Hierarchical Clustering

single $d_{uv,w} = \min(d_{u,w}, d_{v,w})$

complete $d_{uv,w} = \max(d_{u,w}, d_{v,w})$

average $$d_{uv,w} = \frac{1}{N_{uv} N_w} \sum_i \sum_j d_{i,j}$$

Binary variables

a = (1 − 1) matches between object i and j,

b = (1 − 0) mismatches between object i and j,

c = (0 − 1) mismatches between object i and j, and

d = (0 − 0) matches between object i and j.

TEXTBOOK REFERENCES

Johnson & Wichern (2007) *Applied Multivariate Statistical Analysis*

Topic	Chapter	Sections
Matrix manipulation	1 2	All 2.1, 2.2, 2.5, 2.6
Multivariate data	3	3.3-3.6
Multivariate Normal Distribution	4	4.1,4.2,4.4-4.7
Mean Vector Inference	5	5.1, 5.2, 5.4, 5.5
Principal Component Analysis	8	8.1-8.3,8.5
Factor Analysis	9	9.1-9.4
Discriminant Analysis	11	11.1-11.4
Cluster Analysis	12	12.1-12.4