Introduction

Multivariate analysis provides a suite of tools for describing and quantifying the relationship between multiple measured variables.

Often, the primary objective of multivariate analyses is simplification.

Summary of Major Methods

- *Principal component analysis:* Reduce the number of variables, summarize the sources of variation in the data, transform the data into a new data set where the variables are uncorrelated.

- *Factor analysis:* When its not possible to observe the variables of interest directly, measure whats possible and create the variables of interest from the observed data.

- *Discriminant analysis (supervised learning):* Build a rule to predict the class or group id from observed training data.

Summary of Major Methods

- *Cluster analysis (unsupervised learning):* Find similar groups of individuals, or organize the individuals into groups based on their similarity.
- *M(unltivariate)ANOVA*: Infer information about the population means based on the sample means.

Types of Techniques

- Variable-directed: Quantifying the relationships between variables, eg principal component analysis, factor analysis, correlation matrices, regression analysis, canonical correlation analysis.
- Individual-directed: Summarizing relationships that exist between individuals, or experimental units, eg cluster analysis, discriminant analysis, MANOVA.

Matrix Notation

Data (n observations, p variables) has matrix form as follows:

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \ldots \ \mathbf{X}_p]$$

$$= \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}_{n \times p}$$

X_{ij} is the element in the i^{th} row and j^{th} column, that is i^{th} case and j^{th} variable.

For univariate data, the sample *mean* is calculated as

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}, \quad j = 1, \ldots, p.$$

The sample *variance* is calculated as

$$S_j^2 = S_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2, \quad j = 1, \ldots, p.$$

The sample *covariance* is defined as

$$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(X_{ik}-\bar{X}_k), \qquad j,k=1,\ldots,p; j \neq k.$$

The sample *correlation* is defined as $R_{jk} = \frac{S_{jk}}{S_j S_k}$.

Mean Vector, Variance-Covariance/Correlation Matrices

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \dots & S_{pp} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Notation for Projections

A 1-D projection of the data into a vector $\alpha_{\mathbf{p}\times\mathbf{1}}$ takes the form:

$$\begin{aligned}\mathbf{X}\alpha &= [\mathbf{X}_1\alpha\ \mathbf{X}_{\mathbf{2}}\alpha\ \dots\ \mathbf{X}_{\mathbf{n}}\alpha] \\ &= [\alpha_1 X_{11} + \alpha_2 X_{21} + \dots + \alpha_p X_{p1} \quad \dots \\ &\quad\ \alpha_1 X_{1n} + \alpha_2 X_{2n} + \dots + \alpha_p X_{pn}]_{n\times 1}\end{aligned}$$

where $||\alpha|| = \sqrt{\alpha_1^2 + \dots + \alpha_p^2} = \mathbf{1}$. A 2-D projection of the data can be generated by expanding α to $A_{p\times 2} = [\alpha_{\mathbf{1}}\ \ \alpha_{\mathbf{2}}]$ where the columns are orthonormal, $\alpha_{\mathbf{1}}'\alpha_{\mathbf{2}} = \mathbf{0}$. Similarly this notation can be expanded to represent d-D projections.

Distance Measures

Let $\mathbf{A} = (A_1\ X_2\ \ldots A_p)'$ and $\mathbf{B} = (B_1\ B_2\ \ldots B_p)'$ then Euclidean distance is defined as

$$d(\mathbf{A}, \mathbf{B}) = \sqrt{(A_1 - B_1)^2 + \ldots + (A_p - B_p)^2}$$

but *statistical distance* (or Mahalobis distance) is defined as

$$d(\mathbf{A}, \mathbf{B}) = \sqrt{(\mathbf{A} - \mathbf{B})'\mathbf{S}^{-1}(\mathbf{A} - \mathbf{B})}.$$

Generally any distance measure can be defined, but it must satisfy (1) $d(\mathbf{A}, \mathbf{B}) = d(\mathbf{B}, \mathbf{A})$, (2) $d(\mathbf{A}, \mathbf{B}) > 0$, if $\mathbf{A} \neq \mathbf{B}$, (3) $d(\mathbf{A}, \mathbf{B}) = 0$, if $\mathbf{A} = \mathbf{B}$, (4) $d(\mathbf{A}, \mathbf{B}) \leq d(\mathbf{A}, \mathbf{C}) + d(\mathbf{C}, \mathbf{B})$, for any intermediate point $\mathbf{C}$.

Standardized Values and Z-scores

The standardized data matrix is

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

where $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, ..., n; j = 1, ..., p.$

Eigenvectors and Eigenvalues

Any square, symmetric matrix (eg $\mathbf{S}, \mathbf{R}$) can be decomposed or represented in terms of its eigenvalues and eigenvectors.