Multivariate Methods

- *Univariate* methods are used to study a *single* variable either
  - in isolation; or
  - as a response to other factors.

- *Multivariate* methods study the *joint* behavior of two or more (perhaps many) variables.
Possible Goals of a Multivariate Analysis

**Data reduction:** identify one or a few new variables that capture most of the variability in a large data set.

**Sorting and grouping:** develop rules for splitting a large data set into relatively homogeneous groups.

**Dependence:** characterize relationships among variables.

**Prediction:** use dependence to predict.

**Hypothesis testing:** multiple responses in a designed experiment.
Some basics

- Terminology: we measure several variables for each item in a sample; \( p \) variables, \( n \) items.

- Notation: \( x_{j,k} = \) measurement of \( k^{th} \) variable on \( j^{th} \) item.

- Data array:

\[
X = \begin{bmatrix}
  x_{1,1} & x_{1,2} & \cdots & x_{1,k} & \cdots & x_{1,p} \\
  x_{2,1} & x_{2,2} & \cdots & x_{2,k} & \cdots & x_{2,p} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{j,1} & x_{j,2} & \cdots & x_{j,k} & \cdots & x_{j,p} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n,1} & x_{n,2} & \cdots & x_{n,k} & \cdots & x_{n,p}
\end{bmatrix}
\]
Descriptive Statistics

Many multivariate analyses begin with simple summary statistics:

- the row of sample means

\[ \bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{j,k} \]

- the sample variances

\[ s_k^2 = \frac{1}{n} \sum_{j=1}^{n} \left( x_{j,k} - \bar{x}_k \right)^2 \]

and covariances

\[ s_{i,k} = \frac{1}{n} \sum_{j=1}^{n} \left( x_{j,i} - \bar{x}_i \right) \left( x_{j,k} - \bar{x}_k \right). \]
Notes:

- Variances and covariances are often defined with a divisor of $(n - 1)$ instead of $n$—more on that later.

- A variance is the covariance of a variable with itself: $s_k^2 = s_{k,k}$.

- Symmetry: $s_{i,k} = s_{k,i}$.

- Pearson’s product moment correlation

  $$r_{i,k} = \frac{s_{i,k}}{\sqrt{s_{i,i} \times s_{k,k}}}$$

  is dimensionless, and doesn’t depend on choice of divisor.
Arrays of statistics

\[ \bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{bmatrix}, \]

\[ S_n = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p,1} & s_{p,2} & \cdots & s_{p,p} \end{bmatrix}, \]

\[ R = \begin{bmatrix} 1 & r_{1,2} & \cdots & r_{1,p} \\ r_{2,1} & 1 & \cdots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p,1} & r_{p,2} & \cdots & 1 \end{bmatrix}. \]
Graphics

- Make graphs!

- \( p = 2 \): scatter plot—e.g. 1977 baseball salaries and records

  \[
  \text{bb1977} = \text{read.table("JandW/T01-01.dat");} \\
  \text{colnames(bb1977)} = \text{c("Player Payroll", "Won-Lost");} \\
  \text{bb1977;} \\
  \text{plot(bb1977);} \\
  \]

- \( p > 2 \): all pairs of scatter plots—e.g. paper quality

  \[
  \text{paper} = \text{read.table("JandW/T01-02.dat");} \\
  \text{colnames(paper)} = \text{c("Density", "StrengthMD", "StrengthCD");} \\
  \text{paper[1:5, ];} \\
  \text{pairs(paper);} \\
  \]
• When $p >> 2$, too many pairs $(p \times (p - 1)/2)$.

• Two possibilities: stars and Chernoff's faces—e.g. public utilities

```r
util = read.table("JandW/T12-05.dat");
stars(util[1:5, 1:8]);
library(aplpack);
faces(util[ , 1:8]);
```