MULTIVARIATE ANALYSIS: OVERVIEW

by

Ingram Olkin
Allan R. Sampson

February 2001

Department of Statistics
STANFORD UNIVERSITY
Stanford, California
MULTIVARIATE ANALYSIS: OVERVIEW

by

Ingram Olkin
Department of Statistics
Stanford University

Allan R. Sampson
Department of Statistics
University of Pittsburgh

FEBRUARY 2001

This research was supported in part by the
National Science Foundation
grant DMS-0072207 and from the
Alexander von Humboldt Foundation

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
MULTIVARIATE ANALYSIS: OVERVIEW

Ingram Olkin

Department of Statistics, Stanford University, Stanford, CA 94305

and

Allan R. Sampson

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15250

ABSTRACT

Multivariate analysis is conceptualized by tradition as the statistical study of data in which multiple measurements are made on each experimental unit and for which the relationships among multivariate measurements and their structure are important to the study’s understanding. A fairly natural and modern, somewhat overlapping, categorization of multivariate analysis is the following: normal and more general multivariate models and distribution theory; the study and measurement of relationships; probability computations of multidimensional regions; and the study and exploration of data structures and patterns.

The multivariate normal distribution plays a central role in the statistics of multidimensional data in the same way that the univariate normal distribution plays a central role. The Student’s-t generalizes to Hotelling $T^2$ and the chi-square distribution to the Wishart distribution. Newer multivariate distributions have been developed using various approaches to model multivariate data when the multivariate normal distribution does not provide an adequate model. For multidimensional data, relationships among the variables are fundamental to explore. Among the useful techniques to understand and quantify these are multivariate regression analysis and various correlational notions such as partial correlations and canonical correlations. Implementation of multivariate methods often requires computations of complex multidimensional probabilities. Approaches
to these computations include both obtaining lower bounds for the probabilities and using numerical approximation techniques. The exploration of structure and patterns for complex large multivariate data sets is a crucial concern of modern data analysis and data mining.

Multivariate tools that have proven to be useful in this context, include principal components analysis, canonical analysis, factor analysis, path analysis, structural equation methods, clustering, and discriminant analysis.
1. INTRODUCTION

1.1 Study of multiple variables

Multivariate analysis is conceptualized by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding. For instance, in analyzing financial instruments, the relationships among the various characteristics of the instrument are critical. In biopharmaceutical medicine, the patient's multiple responses to a drug need be related to the various measures of toxicity. Some of what falls into the rubric of multivariate analysis parallels traditional univariate analysis; for example, hypothesis tests that compare multiple populations. However, a much larger part of multivariate analysis is unique to it; for example, measuring the strength of relationships among various measurements.

Multivariate analysis, due to the size and complexity of the underlying data sets, requires much computational effort. With the continued and dramatic growth of computational power, multivariate methodology plays an increasingly important role in data analysis, and multivariate techniques, once solely in the realm of theory, are now finding value in application.

1.2 Overview of field

 Whereas there is no standard taxonomy of multivariate analysis, the following presents a fairly natural and modern categorization:

(a) multivariate normal distribution: theory and basic inference;

(b) relationships: regression, correlation, and measures of dependence;

Allan R. Sampson's research is supported by the National Science Foundation, Grant
No. DMS-0072207. Ingram Olkin acknowledges support from the Alexander von Humboldt Foundation.

(c) multivariate models: non-normal multivariate distributions, characterizations, mixture methods, and distributions with given marginals (copulas);

(d) probability computations: inequalities, approximations, and computational methods;

(e) structures and patterns: principal components, canonical analysis, latent structure and causal models, clustering, classification and discrimination, multidimensional scaling, and data mining.

This categorization clearly has overlapping boundaries, and is not all-inclusive.

2. NORMAL THEORY

2.1 Distribution Theory

a) Multivariate normal distribution

In much the same way that the univariate normal distribution is central to statistics, the multivariate normal distribution plays a similar central role. Let \( X = (X_1, \ldots, X_p) \) be a \( p \)-dimensional vector of the multiple random variables measured on an experimental unit. The probability density, \( f(x) \), of \( X \) is multivariate normal, denoted \( N(\mu, \Sigma) \), if

\[
f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-(1/2)(x - \mu)\Sigma^{-1}(x - \mu)'\right\},
\]

for \(-\infty < x_i < \infty, \ i = 1, \ldots, p\), where \( \mu = (\mu_1, \ldots, \mu_p) \) is the \( p \)-dimensional population mean vector \((-\infty < \mu_i < \infty, \ i = 1, \ldots, p)\) and \( \Sigma \) is the \( p \times p \) population (positive definite) variance-covariance matrix. When \( \Sigma \) is singular, the probability density function \( f(x) \) does not exist as noted, but there are other meaningful ways to define the multivariate normal distribution.
The contours of constancy of \( f(x) \) are ellipsoids centered at \( \mu \) with axes determined by the eigenvectors of \( \Sigma \). Any subset of \( X_1, \ldots, X_p \) itself has a multivariate normal distribution and any set of linear combinations has a multivariate normal distribution. Also the conditional probability distribution of any subset given another subset has a multivariate normal distribution.

Other multivariate normal distributional properties have been extensively studied. For example, if \( \Sigma = (\sigma_{ij}) \) contains all nonnegative elements, then \( P\{X_1 \leq x_1, \ldots, X_p \leq x_p\} \geq P\{X_1 \leq x_1\} \times \cdots \times P\{X_p \leq x_p\} \) for all \( x_1, \ldots, x_p \), and the same inequality holds if "\( \leq \)" inside all the probability statements are replaced by "\( \geq \)". The first inequality holds for any covariance matrix when, for all \( i \), \( |X_i| \) is used instead of \( X_i \). Moreover, if \( \sigma_{ij} \geq 0 \), for all \( i, j \), then the correlation is nonnegative between any two increasing functions of \( |X_1|, \ldots, |X_p| \); for example, in this setting, the correlation between \( \sum_{i=1}^{p} X_i^2 \) and \( \prod_{i=1}^{p} |X_i| \) is nonnegative.

b) Sample statistics and their properties

Suppose that a random sample \( X_1, \ldots, X_N \), of size \( N \), is observed from a multivariate normal distribution. The natural estimators of \( \mu \) and \( \Sigma \) are, respectively, the sample mean vector \( \bar{X} \) and the sample variance-covariance matrix \( S = \sum_{i=1}^{N} (x_i - \bar{x})'(x_i - \bar{x})/(N-1) \).

These estimators possess many of the properties of the univariate sample mean and variance. They are independent and unbiased; furthermore, \( \bar{X} \) and \( [(N-1)/N]S \) are maximum likelihood estimators. The sampling distribution of \( \bar{X} \) is \( N(\mu, \Sigma/N) \), and the sampling distribution of \( (N-1)S \) is called the Wishart distribution with parameter \( \Sigma \) and \( N-1 \) degrees of freedom, as obtained by John Wishart in 1928.

For the univariate normal distribution, when the variance \( \sigma^2 \) is known, the mean \( \bar{x} \) is an optimal estimator of the population mean \( \mu \) in any reasonable sense. It was shown in 1956 by Charles Stein that when \( p > 2 \), and the underlying multivariate distribution is \( N(\mu, I) \), one can improve upon \( \bar{X} \) as an estimator of \( \mu \) in terms of its expected precision, because surprisingly, even with the identity matrix \( I \) being the population variance-covariance matrix, there is information in \( S \) to improve estimation of \( \mu \).

Authoritative references to the multivariate normal distribution and its properties
include Anderson (1984), Muirhead (1982), and Tong (1990).

c) Characteristic Roots

When the sample covariance matrix $S$ has a Wishart distribution with the identity population covariance matrix, many invariant tests are functions of the characteristic roots of $S$. This led to the study of the joint distribution of the roots, which was obtained simultaneously by several authors in 1939. Characteristic roots of a sample covariance matrix also play a role in physics and other fields.

The basic distribution theory for the case of a general diagonal covariance matrix was developed by Alan James in 1956 and requires expansion in terms of zonal polynomials. For details of the noncentral distribution of the characteristic roots and zonal polynomials see Muirhead (1982), and Mathai, Provost, and Hayakawa (1995).

2.2 Inference

Inference concerning $\mu$ when $\Sigma$ is known is based, in part, upon the Mahalanobis distance $N(\bar{X} - \mu)\Sigma^{-1}(\bar{X} - \mu)'$ which has a $\chi^2_N$ distribution when $X_1, \ldots X_N$ is a random sample from $N(\mu, \Sigma)$. When $\Sigma$ is not known, inference about $\mu$ utilizes the Mahalanobis distance with $\Sigma$ replaced by its estimator $S$. The distribution of the quantity $N(\bar{X} - \mu)S^{-1}(\bar{X} - \mu)'$ was derived by Harold Hotelling in 1931 and is called Hotelling’s $T^2$. He showed that $(N - p)T^2/(N - 1)p$ has a standard $F$-distribution, $F_{p,N-p}$.

To test, at level $\alpha$, the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_A : \mu \neq \mu_0$ based upon multivariate normal data with unknown $\Sigma$, one rejects $H_0$ if $N(\bar{X} - \mu_0)S^{-1}(\bar{X} - \mu_0)'$ exceeds $(N - 1)p/(N - p)F^\alpha_{p,N-p}$, where $F^\alpha_{p,N-p}$ denotes the upper $\alpha$-critical value of $F_{p,N-p}$. Elliptical confidence intervals for $\mu$ can be constructed as $\{\mu : N(\bar{X} - \mu)S^{-1}(\bar{X} - \mu)' \leq ((N - 1)p/(N - p)) \times F^\alpha_{p,N-p}\}$. Useful rectangular confidence intervals for $\mu$ are more difficult to obtain and typically rely on suitable probability inequalities. (see Hochberg and Tamhane (1987)). For the univariate normal distribution, one-sided hypothesis testing straightforwardly follows from two-sided testing. However, various one-sided tests for the multivariate case are much more complicated (see Perlman and Wu
A particular class of techniques that are useful in this case are the techniques of order-restricted inference, e.g., Robertson, Wright and Dykstra (1988)).

The comparison of two multivariate normal population means, with common but unknown variance-covariance matrix can be based upon a two-sample version of Hotelling’s $T^2$ statistic.

2.3 Linear Models: Multivariate Analysis of Variance (MANOVA)

To compare $k (> 2)$ multivariate normal population means (with common unknown variance-covariance matrix), one needs to utilize generalizations of Hotelling’s $T^2$ statistic, similar to the relationship of the $F$-statistic for multiple populations to the two-sample $t$ for univariate data.

To understand more generally multivariate analysis of variance, consider for the univariate setting the analysis-of-variance (ANOVA) for a given hypothesis.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>F-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>$SS_H$</td>
<td>$d_H$</td>
<td>$F = (SS_H/d_H)/(SS_E/d_E)$</td>
</tr>
<tr>
<td>Error</td>
<td>$SS_E$</td>
<td>$d_E$</td>
<td></td>
</tr>
</tbody>
</table>

The null hypothesis reference distribution for $F$ is $F_{d_H, d_E}$ and the alternative distribution is a non-central $F$-distribution; these result from $SS_H$ and $SS_E$ being independent random variables with $SS_E \sim \sigma^2 \chi^2_{d_E}$ and under $H_0$, $SS_H \sim \sigma^2 \chi^2_{d_H}$. For example, in a one-way layout, $SS_H$ is the sum-of-squares for treatment, and $SS_E$ is the usual sum-of-squares for error.

The general structure of an ANOVA table for multivariate data of dimension $p$ is parallel in structure; $S_H$ becomes $H$, a $p \times p$ matrix of sum-of-squares and cross-products, and $SS_E$ becomes $E$, a $p \times p$ matrix related to the variance-covariance matrix of the multivariate error vector. The random matrices $H$ and $E$ are independent; $E$ has a $p$-dimensional Wishart distribution with $d_E$ degrees of freedom, and under $H_0$, $H$ has a
p-dimensional Wishart distribution with degrees of freedom $d_H$.

However, now the ratio of sums of squares is translated to the eigenvalues of $HE^{-1}$. Because there is no uniformly best function of these eigenvalues, various alternative functions of the product, sum, or maximum of the eigenvalues have been proposed. The most well known statistics are the Wilks lambda, Hotelling-Lawley trace, Pillai trace, and Roy maximum, and are provided in standard MANOVA statistical packages.

3. RELATIONSHIPS

3.1 Regression

Regression analysis is designed to predict a measure $Y$ based on concomitant variables $X = (X_1, \ldots, X_p)$. It was historically traditional to assume that $(Y, X)$ have a joint normal distribution with zero means and $(p + 1) \times (p + 1)$ covariance matrix of $(Y, X)$: $\text{Var}(Y) = \sigma_{00}, \text{Covar}(Y, X) = \Sigma_{12}, \text{Var}(X) = \Sigma_{22}$. Then the conditional mean of $Y$ given $X = x$ is $\Sigma_{12}\Sigma_{22}^{-1} x$, which is a linear function of the form $\beta_1 x_1 + \cdots + \beta_p x_p$; the conditional variance is $\sigma_{00} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$, which is independent of $x$, a fact that is critical in the analysis.

Although the above description involves the multivariate normal distribution, the subject of regression is usually treated more simply as ordinary linear regression involving least squares. See the entry on Regression for a history and further details, both for the multivariate normal setting, as well as other settings.

The more general regression model consists of a $p \times n$ matrix $Y$ in which the $n$ columns are independently distributed, each having an unknown $p \times p$ covariance matrix $\Sigma$, and the expected value of $Y$ is a function of unknown parameters, namely, $E \, Y = X_1 B X_2$, where $X_1$ and $X_2$ are $p \times q$ and $r \times n$ matrices, respectively, and $B$ is a $q \times r$ matrix of unknown parameters which is to be estimated. This model comprises what is called the multivariate general linear model (see Anderson (1984) for details).
3.2 Correlation hierarchy: partial, multiple, canonical correlations

If \( \mathbf{X} \) is a \( p \)-dimensional random vector with covariances \( \sigma_{ij} \), the population correlation coefficient between \( X_i \) and \( X_j \) is defined as \( \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}} \), with the sample analogue correspondingly defined. Alternative names are product moment correlation or Pearson's correlation coefficient; it was Karl Pearson who introduced the sample correlation in 1896. Because \( -1 \leq \rho_{ij} \leq 1 \) and \( \rho_{ij} = \pm 1 \) only when \( X_i \) and \( X_j \) are perfectly linearly related, the correlation has become a common measure of linear dependence.

For normally distributed data the distribution of the sample correlation was obtained in 1915 by R. A. Fisher. The test statistic \( \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \) can be used to test that the population correlation is zero; under this hypothesis, this statistic has a Student's \( t \) distribution with \( n-2 \) degrees of freedom. Fisher later obtained a transformation to \( z = \tan h^{-1}(r) \), which he showed to have an approximate normal distribution with mean \( \tan h^{-1}(\rho) \) and variance \( (n-3)^{-1} \). Consequently, the \( z \) transformation provides confidence intervals for \( \rho \). The definitive analysis of the distribution of \( r \) and \( z \) is given by Hotelling (1953).

The population multiple correlation coefficient arises when \( X_2, \ldots, X_p \) are used to predict \( X_1 \), and is defined as the population correlation between \( X_1 \) and the best linear predictor of \( X_1 \) based on \( X_2, \ldots, X_p \). A similar definition holds for the sample multiple correlation. The population partial correlation is the correlation between a pair of variables, say, \( X_1 \) and \( X_2 \), conditional on the remainder being fixed, that is, \( X_3, \ldots, X_p \) being fixed, with a similar definition holding for the sample. See Anderson (1984) for results concerning the distributions of the sample multiple correlation coefficient and the sample partial correlation coefficient, as well as related techniques for statistical inference, in the case of normally distributed data. Canonical correlations are used to measure the strength of relationship between two sets of variables and are described more fully in Section (6.2).
3.3 Measures of dependence

Although Pearson's correlation coefficient is a natural way to measure dependence between random variables $X$ and $Y$ having a bivariate normal distribution, it can be less than meaningful for non-normal bivariate distributions. Numerous other measures of dependence between arbitrary random variables have been introduced, as have techniques for estimating each from data. Popular and useful population measures for continuous random variables include Kendall's $\tau$ ($= 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x, y)f(x, y)dxdy - 1$) and Spearman's rho ($= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x, y) - F_x(x)F_y(y)) f_x(x)f_y(y)dxdy$), where $F(x, y), F_x(x), F_y(y)$ are, respectively, the bivariate distribution function of $X$ and $Y$, and the two marginal distribution functions of $X, Y$; and where $f(x, y), f_x(x)$, and $f_y(y)$ are their corresponding density functions. Kendall's $\tau$ can be written as $4 E(F(X, Y)) - 1$, where the expectation is with respect to the joint distribution of $X, Y$ and Spearman's rho can be written as $12 E[F(X, Y) - F_x(X)F_y(Y)]$, where the expectation is with respect to independent $X$ and $Y$, with distributions $f_x(x)$ and $f_y(y)$. The sample estimates of Kendall's $\tau$ and Spearman's rho are given in Agresti (1984).

Other more theoretical approaches to measuring dependence are based upon maximizing $\rho(a(X), b(Y))$ over certain classes of functions $a(x)$ and $b(y)$. Renyi's sup-correlation is the case for essentially general functions, and Kimeldorf and Sampson's (1978) monotone correlation maximizes only over monotone functions.

The preceding measures of dependence are also applicable to bivariate discrete random variables which can be measured on an ordinal scale, i.e., an ordinal cross-classification. For nonordinal, i.e., nominal cross-classifications, there are other appropriate measures of dependence (see Agresti (1990), or Goodman and Kruskal (1979) which is a classical reference source).

3.4 Canonical decomposition

As a means of relating a bivariate distribution to sums of independent-like components, the canonical decomposition (see Lancaster (1969)) is useful. With the assumption of certain technical conditions, the joint density function of continuous random variables
can be written as \( f(x,y) = \{1 + \sum_{i=1}^{\infty} \rho_i a_i(x)b_i(y)\}f_x(x)f_y(y) \), where the coefficients \( \rho_i \) are called the generalized canonical correlations, and \( a_i(x) \) and \( b_i(y) \), \( i = 1, \ldots \), are well-behaved functions. Furthermore, when \( X,Y \) are jointly discrete, an equivalent canonical decomposition exists and is related to the spectral decomposition of matrix theory. This decomposition and modified versions have been employed for the analysis of contingency tables and in correspondence analysis. For a more general discussion, see Agresti (1990) and Greenacre (1994).

4. METHODS FOR GENERATING MULTIVARIATE MODELS

Many newer multivariate distributions have been developed to model data where the multivariate normal distribution does not provide an adequate model. Multivariate distributional modeling is inherently substantially more difficult in that both marginal distributions and joint dependence structure need to be taken into account. Rather straightforward univariate properties or characterizations often have many multivariate versions; however, numerous multivariate concepts have no univariate analogues. Some of the more commonly used approaches to developing models are given in the following.

4.1 Characterizing Properties

One common approach is to consider multivariate extensions of a specific univariate property. For instance, symmetry of the univariate distribution, that is, \( f(x) = f(-x) \), can lead to the family of reflection symmetric or multivariate spherically distributions. Here \( f(x_1, \ldots, x_p) = g(|x_1|, \ldots, |x_p|) \) or \( f(x_1, \ldots, x_p) = g(x_1^2 + \cdots + x_p^2) \) for a suitable function \( g \). The latter family, especially when incorporating covariates, naturally leads to the class of elliptically contoured multivariate distributions. (e.g., Fang and Anderson (1990)).

The lack of memory property, which characterizes the univariate exponential distribution, has been generalized to the multivariate case, and this characterization yields the
Marshall-Olkin multivariate exponential distribution. Because this distribution has both a singular and a continuous part, other purely continuous multivariate distributions with exponential marginals have been developed.

Other properties that have been used to develop multivariate distributions and independence of certain statistics, identical distribution of statistics, closure properties, limiting distributions, differential equation methods, series expansions, and conditioning methods. In some instances physical models, such as shock models, provide a method for extending univariate distributions to the multivariate case.

4.2 Mixture Methods

Mixture methods have provided successful extensions for a large class of models, especially for non-negative random variables. Several examples will suffice to indicate the underlying procedure.

Suppose $U, V, W$ are independent random variables; then bivariate random variables can be defined by each of the following:

(i) $X_1 = U + V, \ X_2 = U + W,$

(ii) $X_1 = \min(U, V), \ X_2 = \min(U, W),$ ,

(iii) $X_1 = U/V, \ X_2 = U/W.$

A bivariate binomial distribution can be generated using (i), when $U, V, W$ are each respectively binomially distributed; a bivariate exponential distribution can be generated using (ii) when $U, V, W$ are exponential; and a bivariate Pareto or $F$ distribution can be generated using (iii) when $U, V, W$ each have a gamma distribution. By taking limits, these families can be further broadened. For a review of the relationship among bivariate distributions see Marshall and Olkin (1985), and for a comprehensive treatment of bivariate distributions, see Hutchinson and Lai (1990).

One of the inherent problems with multivariate versions of mixture methods (as well as some other techniques for generating multivariate distributions) is that the number of parameters so introduced can be large. For instance, one general extension of (i) to $p$
variables requires \(2^p - 1\) summands, each one of which corresponds to a parameter. This implies the need for large sample sizes in order to estimate the parameters for such a multivariate distribution.

### 4.3 Copulas

Another approach to providing non-normal multivariate distributional models is based on copulas or uniform representations. We provide details for bivariate random variables; the more general case is discussed in Nelsen (1999) and Joe (1997). This approach first requires specification of the univariate marginal distribution functions \(F_1(x), F_2(y)\), and then the joint distribution is \(F(x, y) \equiv U(F_1(x), F_2(y))\), where \(U(s, t)\) is a two-dimensional copula. There are a number of technical conditions that the function \(U(s, t)\) must satisfy, but the basic idea is that \(U(s, t)\) is essentially the distribution function of a two-dimensional random vector whose every component marginally has a uniform distribution on \([0, 1]\).

There are many different families of copulas; for the bivariate case, examples of copulas include the standard normal copula,

\[
(a) \quad U_N(s, t) = (1 - \rho^2)^{-1/2} \exp[-\{\rho^2(u^2 + v^2) - 2\rho uv\} / (2(1 - \rho^2))] ,
\]

where \(u = \Phi^{-1}(s)\), \(v = \Phi^{-1}(t)\) and \(\Phi^{-1}\) is inverse standard normal distribution function; and the Farlie-Gumbel-Morgenstern copula,

\[
(b) \quad U_{FGM}(s, t) = 1 + \theta(2s - 1)(2t - 1).
\]

Based upon copulas like these, different multivariate distributions can be generated. For instance, the standardized bivariate normal distribution function with correlation \(\rho\) can be expressed as \(U_N(\Phi(x), \Phi(y))\), whereas \(U_N(1 - e^{-x}, 1 - e^{-y})\) has normal-like dependence structure but standard exponential distributions as its marginals. A fairly standard technique for creating copulas is to begin with any joint distribution \(F(x, y)\) and then compute \(F(F_1^{-1}(x), F_2^{-1}(y))\), which under nonstringent technical conditions will be a copula.

Any bivariate copula \(U(s, t)\) with \(0 \leq s, t \leq 1\) satisfies \(U^-(s, t) \equiv \min(s, t) \leq U(s, t) \leq \max(s + t - 1, 0) \equiv U^+(s, t)\).
Both $U^{-}(s_1, s_2)$ and $U^{+}(s_1, s_2)$ themselves are copulas, and are called, respectively, the lower and upper Fréchet-Hoeffding bounds.

The probabilistic theory concerning copulas has been fairly extensively developed, but the applications to data analysis are much more limited.

4.4. Other multivariate models

For multivariate categorical data, a popular class of models are the log-linear models, where essentially $\log(f(x_1, \ldots, x_p))$ is written as a sum of terms, each term depending on subsets of $x_1, \ldots, x_p$, where $f(x_1, \ldots, x_p)$ is the joint probability mass function. For example, the log linear model $\log f(x_1, \ldots, x_p) = \mu + \lambda_1 + \cdots + \lambda_p$ corresponds to independent random variables $X_1, \ldots, X_p$ (e.g., Agresti (1990)). Related models are certain multivariate logistic models (see Glonek and McCullagh (1995)) which include covariates in the joint p.m.f. by means of a multivariate logistic transformation, and yield marginal distributions which are univariate logistically distributed.

Other models than the Marshall-Olkin multivariate exponential distribution have been employed to model dependent life times. One such model is the bivariate frailty model (also related to the archimedean copula) which allows an unobservable frailty to cause dependence in survival times which marginally can have a proportional hazards model (see Cox and Oakes (1984)). Other multivariate survival distributions and multivariate extreme value distributions have also been developed.
5. MULTIVARIATE PROBABILITY

COMPUTATIONAL TECHNIQUES

5.1 Multivariate bounds for probabilities

Inequalities or bounds are important tools in simultaneous inference and for approximating probabilities. We consider three basic types of inequalities, noting that for each there are numerous generalizations and extensions.

Perhaps the oldest method is based on Boole inequalities which provide relations between probabilities of events. Suppose that we know the probabilities of several events, such as the probability of heart failure (A) and the probability of lung failure (B), that is, we have estimates of $P\{A\}$ and $P\{B\}$. Then the probability of the joint event is bounded

$$\max [0, P\{A\} + P\{B\} - 1] \leq P\{A \text{ and } B\} \leq \min [P\{A\}, P\{B\}] .$$

The number of inequalities increases as the number of events increase. The Fréchet-Hoeffding bounds can be viewed as extensions of Boole's inequality. In a statistical context, the lower bound of the Boole inequalities are called Bonferroni inequalities, and are often used when dealing with multiple comparisons. For a set of hypotheses $H_1, \ldots, H_m$, then the probability of rejecting any of them

$$P\{\text{Reject } H_1 \text{ or } \cdots \text{ or Reject } H_m\} \leq P\{ \text{ Reject } H_1\} + \cdots + P\{ \text{Reject } H_m\} .$$

The choice of $P\{\text{Reject } H_i\} = \alpha/m$ controls the error rate of rejecting any of the hypotheses to be at most $\alpha$.

These bounds have many extensions, for example, when the probabilities of pairs $P\{\text{Reject } H_i \text{ or Reject } H_j\}$ are known. [For a general discussion of multiple comparisons see Hochberg and Tamhane (1987), and for Bonferroni bounds see Galambos and Simonelli (1996).]

A second set of inequalities are called Chebyshev inequalities. In the univariate setting, given the mean $\mu$ and variance $\sigma^2$, the probability of the symmetric two-sided tail events
is restricted regardless of the underlying distribution, that is, the classical Bienaymé-
Chebyshev inequality asserts that

\[ P\{|X - \mu| \geq \sqrt{k\sigma}\} \leq 1/k. \]

For bivariate random variables \(X_1, X_2\) with \(EX_i = \mu_i, \text{Var}(X_i) = \sigma_i^2\) and correlation \(\rho\),

\[ P\{|X_1 - \mu_1| \geq \sqrt{k_1 \sigma_1}, |X_2 - \mu_2| \geq \sqrt{k_2 \sigma_2}\} \leq \frac{k_1 + k_2 + \sqrt{(k_1 + k_2)^2 - 4\rho^2 k_1 k_2}}{2k_1 k_2}. \]

Further knowledge of moments yields still tighter inequalities.

Another class of inequalities are multiplicative inequalities that hold for certain classes
of random variables. The general form of these inequalities is of the form

\[ P\{X_1 \in A_1, \ldots, X_m \in A_m\} \geq \prod_{i=1}^m P\{X_i \in A_i\}. \]

For instance, this inequality holds for all intervals \(A_i = (t_i, \infty)\) when the random variables
are positively upper orthant dependent. In particular, when \(X_1, \ldots, X_m\) are independent
normal variates with respective means \(\mu_1, \ldots, \mu_m\), and common variance \(\sigma^2\), then

\[ P\{\cap |X_i - \mu_i| \geq k_i \delta\} \geq \prod_{i=1}^m P\{|X_i - \mu_i| \geq k_i \delta\}, \]

where \(\delta^2\) is an independent estimate of \(\sigma^2\). This inequality yields a bound for the joint
probability of a simple multivariate \(t\)-distribution in terms of probabilities of univariate \(t\)-distributions.

When \(X_1, \ldots, X_m\) are correlated normal variates with a correlation matrix \(R = (\rho_{ij})\)
with \(\rho_{ij} \geq 0\), then the one-sided orthant is bounded

\[ P\{\cap (X_i \leq a_i)\} \geq \prod_{i=1}^m P\{X_i \leq a_i\}. \]

The two-sided version of the preceding holds for \(|X_i|\) without any non-negativity restrictions on the correlations.

For discussions of multivariate inequalities see Dharmadhikari and Joag-dev (1988)
and Tong (1980).
5.2 Approximations of probabilities

There are a number of different numerical and stochastic techniques that are in use to obtain numerical approximations to multivariate probability statements. These techniques are useful in a variety of circumstances, for example, computing p-values for complicated multivariate data, obtaining power curves for multivariate tests, and computing multivariate posterior distributions.

Basic techniques for multivariate probability approximations rely on standard numerical integration ideas where \( P(x \in A) = \int f(x) \, dx \) is approximated by \( \sum_{\{x_i\}} w_i f(x_i) \) where \( w_i \) are weights and \( \{x_i\} \) is a finite set of points specified in \( A \). Elementary Monte-Carlo integration techniques involve random sampling \( N \) points \( x_i, \ i = 1, \ldots, N \) according to \( f(x) \) and estimating \( P(A) \) by \( N^{-1} \sum_{i=1}^{N} I(x_i \in A) \), where \( I(x_i \in A) = 1 \), if \( x_i \in A \), and \( = 0 \), otherwise. Note that by the strong law of large numbers \( N^{-1} \sum_{i=1}^{N} I(x_i \in A) \) converges almost surely to \( P(A) \) as \( N \to \infty \). Monte-Carlo integration can be modified to estimate \( E(h) = \int h(x) \, f(x) \, dx \) by \( N^{-1} \sum_{i=1}^{N} h(x_i) \), where again \( x_1, \ldots, x_N \) are sampled from \( f(x) \). Some methods for computing, in particular, multivariate normal probabilities first transform the random variables, and rewrite \( P(A) = \int_{||z||=1} F(z) \, dz \) where \( F(z) \) is a function defined by \( \Sigma \) and the region \( A \). Then \( P(A) \) is estimated by \( N^{-1} \sum_{i=1}^{N} F(z_i) \) where \( z_i \) are randomly chosen from the surface of a sphere in \( p \)-dimensional space.

Monte-Carlo importance sampling techniques suppose that there is a convenient density \( g(y) \) whose support includes the support of the density \( f(x) \). Then points \( y_1, \ldots, y_N \) are random sampled according to \( g(y) \) and \( P(A) \) is estimated by \( N^{-1} \sum_{i=1}^{N} I(y_i \in A)(f(y_i)/g(y_i)) \). The idea is to choose \( g(y) \) to speed the convergence of the random sum to \( P(A) \), its true mean. For standard Monte-Carlo methods, the usual estimates of the standard error of estimation provide error estimates for the probability approximation. See Davis and Rabinowitz (1984) for a comprehensive and general discussion of numerical integration techniques.

More recently Markov chain Monte-Carlo methods have been developed and have proven to be particularly useful for high dimensional integration. The common aspects
of these types of techniques is the definition of a Markov chain \( x_1, x_2, \ldots \) which converges to a stationary Markov chain where \( x_i \) has marginal distribution \( f(x) \). Again, \( f \cdots f h(x)f(x)dx \) is approximated by \( (N - m + 1)^{-1} \sum_{i=m}^{N} h(x_i)f(x_i) \) where \( x_m, \ldots, x_N \) is a suitably chosen sequence of elements generated by the Markov chain. In certain settings where particular conditional distributions can be specified among the components of \( x \), the Gibbs sampling algorithm is a Markov chain type method which has proven to be quite useful. For a discussion of these and other related algorithms including the Metropolis algorithm, see Evans and Swartz (1995).

6. STRUCTURES AND PATTERNS

6.1 Principal components

A correlation matrix represents the interdependencies among the \( p \) measures, which may be likened to a connected network.

The removal of one of two closely connected variables (that is, highly correlated) takes no account of how these variables are connected to the remaining measures.

Principal components introduced by Harold Hotelling in 1933, is a methodology that transforms the original correlated

standardized variables \( z_1, \ldots, z_p \) into new variables (called principal components) \( z_1^*, \ldots, z_p^* \), which are uncorrelated. These new variables are linear combinations of the original variables. The variance of the standardized \( z_i \) is 1, and denote the variance of \( z_i^* \) by

\[ \tau_i^2 \]

; then the total variance \( p = \sum \tau_i^2 \) remains fixed for the two systems of variables.

By ordering the variances \( \tau_i^2 \geq \tau_{(2)} \geq \cdots \geq \tau_{(p)} \), and examining the cumulative proportions

\[ \tau_{(1)}/p, (\tau_{(1)} + \tau_{(2)})/p, \ldots, (\tau_{(1)} + \tau_{(p)})/p \]
\[ \cdot + \tau^2_{(p)} \] / p = 1, we obtain what is often called the “proportion of variance explained by the principal components.”

In this way we may obtain a subset of principal components that explain most of the variance, thereby providing a parsimonious explanation of the data.

In effect, principal components provide a method of data reduction. The challenge in using this approach is deciding how many principal components are reasonably required and how to interpret each component.

We now provide the details for how principal components may be obtained. Consider a linear combination \( z_1^* = a_{11}z_1 + \cdots + a_{1p}z_p \) with variance \( \tau^2_{(1)} = \Sigma_{a,p} a_{1\alpha \rho \beta} a_{1\beta} \). The optimization problem is to maximize \( \tau^2_{(1)} \) subject to \( \Sigma a_{1j}^2 = 1 \). This maximization yields the vector \((a_{11}, \ldots, a_{1p})\). Now we choose a second linear combination \( z_2^* = a_{21}z_1 + \cdots + a_{2p}z_p \), with \( \Sigma a_{2j}^2 = 1 \). We maximize Var \((z_2^*) = \Sigma_{a,\rho} a_{2\alpha \rho \beta} a_{2\beta} \) subject to \( z_1^* \) and \( z_2^* \) being uncorrelated. At stage \( j \) we maximize the variance subject to the condition that suitably defined \( z_j^* \) be uncorrelated with \( z_1^*, \ldots, z_{j-1}^* \), for \( j = 2, 3, \ldots, p \).

For the sample we substitute the sample correlation matrix for the population correlation matrix.

### 6.2 Canonical variables

As noted, principal components provides a procedure for the reduction of a large number of variables into a smaller number of new variables. When the variables fall into two natural subsets, as for example production and sales variables, or measurements on two siblings, then canonical analysis provides another exploratory data reduction procedure. If the two sets are labeled \( X_1, \ldots, X_p \) and \( Y_1, \ldots, Y_q \) \((p \leq q)\), canonical analysis attempts to simplify the correlational structure consisting of \((p + q)(p + q - 1)/2\) correlations to a simpler correlational structure of only \( p \) correlations.

This reduction is accomplished by forming linear combinations \( U_1 \) and \( V_1 \) of the \( X_i \)'s and \( Y_j \)'s, respectively. The coefficients are chosen so that \( \text{Var}(U_1) = \text{Var}(V_1) = 1 \) and the correlation, \( \rho_1 \), is maximized. A second pair \( U_2, V_2 \) of linear combinations of the \( X \)'s and \( Y \)'s is chosen to have unit variance, to be uncorrelated with \( U_1, V_1 \) and to have maximum correlation \( \rho_2 \). This procedure is continued to provide pairs \( (U_i, V_i) \), \( i = 1, \ldots, p \), for
which

\[ \text{Var}(U_i) = \text{Var}(V_i) = 1, \]
\[ \rho(U_i, U_j) = \rho(V_i, V_j) = \rho(U_i, V_j) = 0, \ i \neq j, \]
\[ \rho(U_i, V_i) = \rho_i, \ i = 1, \ldots, p. \]

The new variables are called canonical variables, and the correlations \( \rho_i \) are called the population canonical correlations.

This procedure is equally applicable to sample data. Often the first few sample canonical correlations are large compared to the remaining ones, in which case the first few canonical variables provides a data reduction procedure, which captures more of the structure between the two sets of variables.

### 6.3 Latent structure and causal models

Latent structure models refers a set of models that attempts to capture an understanding of causality, and hence are sometimes referred to as causal models. The term is not well-defined and at its broadest includes factor analysis, path analysis, structural equation models, correspondence analysis, loglinear models, and multivariate graphical analysis. One feature of some of these models is the description of a set of observables in terms of some underlying unobservable random quantities. Depending on the context, these unobservable variables have been called latent variables, factors, or manifest variables.

Graphical models often provide a visual understanding of relationships. Their origin arose in a number of scientific areas: as path analysis from genetic considerations, as interactions of groups of particles in physics, as interactions in multi-way contingency tables. They are particularly useful in multivariate models because variables are related in a variety of different ways, as for example, conditionally, longitudinally, or directly. Graphical models can be used in conjunction with specific models, for example, factor analysis, and structural equation models. For an exposition of graphical models see Lautitzen (1996) or Edwards (1995), and for graphical models in multivariate analysis see Cox and Wermuth (1996).
a) Factor analysis

Factor analysis is one of the oldest structural models having been developed by Spearman in 1904. He tried to explain the relations (correlations) among a group of test scores, and suggested that these scores could be generated by a model with a single common factor, which he called "intelligence," plus a unique factor for each test.

In this context, $p$ test scores generate $p(p - 1)/2$ correlations, whereas the factor model contains $p + 1$ parameters. For example, with $p = 10$, we explain 45 correlations with only 11 parameters.

In a more general context let $x_1, \ldots, x_p$ be $p$ observable measures, each of which is linearly expressed in terms of $m$ unobservable common factors and a unique factor. The objective in part is to reproduce the observed correlation network of the $x$'s by the correlations that would ensue with fewer factors.

Factor analysis has been used in two data analytic contexts. When there is a prior theoretical construct that specifies a model, factor analysis is used in a confirmatory manner designed to confirm or negate the hypothesized structure. Where there is no theoretical construct, factor analysis may be used to try to discover a structure, in which case the analysis is called exploratory.

Because factor analysis does not lead to a unique solution, various alternative approaches (called rotating methods) have been designed to provide interpretability.

b) Path Analysis

The origins of path analysis is attributed to Sewall Wright, who from 1918 to 1934 developed a method for studying the direct and indirect effects of variables. Herein, some of the variables are often taken as causes, whereas others are taken as effects. However, because the linear relations exhibit only correlations, the view of causality remains hypothetical.

Path diagrams are typical graphical devices to show the direct and indirect variables. For example, in a study on the relation between performance and satisfaction,
let $x_1$ be achievement motivation, $x_2$ be verbal intelligence, $y_1$ be performance, and $y_2$ be job satisfaction. There are several path diagrams that could exhibit these connections.

![Path Diagrams](image)

In model A, variation in $x_1$ and $x_2$ affect both outcomes $y_1$ and $y_2$, whereas in model B, verbal intelligence affects performance directly, and affects job satisfaction through achievement motivation. Graphical visualization is helpful when the number of variables is not large, but becomes too cumbersome for larger systems.

Each path represents a linear regression. An often difficult part of the analysis is the determination of whether the parameters in the system are identifiable and can be estimated.

c) **Linear structural equations (LISREL)**

Structural equation models, or econometric models, were developed early on to provide explanations of economic measures. Variables whose variability is generated outside the model are called *exogenous* and variables explained by exogenous variables or other variables in the model are called *endogenous*.

A standard basic model is for vectors $x = (x_1, \ldots, x_q)$ and $y = (y_1, \ldots, y_p)$ of observable variables

$$
y = \eta \Lambda_1 + \epsilon, \quad x = \xi \Lambda_2 + \delta,
$$

where $\Lambda_1, \Lambda_2$ are, respectively, $m \times p$ and $q \times n$ matrices of regression coefficients; $\eta$ and $\xi$ are vectors of unobservable dependent variables; $\epsilon$ and $\delta$ are error vectors.
The structural equation that connects the latent exogenous and endogenous variables is

\[ \eta \Delta = \xi \Gamma + \zeta, \]

where \( \eta \) is of dimension \( m \), \( \xi \) is of dimension \( n \), \( \Delta \) is an \( m \times m \) regression matrix of the effects of endogenous on endogenous variables, \( \Gamma \) is an \( m \times n \) regression matrix of the effects of exogenous on endogenous variables, and \( \zeta \) is a vector of residuals.

The LISREL computer programs, due in large part to Karl Jöreskog in the early 1970's, was designed to provide estimates of the parameters. The current LISREL programs and other alternative programs, now permit analysis of a broad spectrum of models.

d) Cross-classification models

Latent structure models in the context of contingency tables were introduced by Paul Lazarsfeld in 1950 to analyze causal hypotheses relating a set of manifest (observable) variables to a set of latent factors. In effect we have a contingency table with cell frequencies that provide an observed joint distribution. The objective is to represent, within acceptable random fluctuation, a model that yields a hypothetical joint distribution.

For a fuller explanation of variants of such models, see Lazarsfeld and Henry (1968).

6.4 Clustering

Often subsets of individuals are grouped by a known characteristic, as for example, by sex, by race, by socio-economic status. In other instances prespecified grouping characteristics may not be available. This is analogous to thinking of a galaxy of stars to be separated into groups. The methodology of clustering may be thought of as a data reduction procedure,

where the reduction is to group individuals.

This is in contrast to principal components or canonical analysis, in which the reduction is based upon combining variables.
Alternatively, we may think of clustering as creating a classification or typology of individuals or items.

For example, we may think of creating regions based on a set of demographic variables, or a set of products based on how they are used.

Suppose that the data consists of \( p \) measurements taken on each of \( n \) individuals (objects). For each pair of individuals, we form a “distance” measure, where “distance” is flexibly defined as, for example, Euclidean distance, correlation, similarity measures, and dissimilarity measures. Thus, we create an \( n \times n \) matrix of “distances.”

Based upon this matrix, there are a variety of algorithms that govern the creation of subgroups, in part because there is no universal definition of a cluster, nor a model that defines a cluster.

Intuitively, we think of a cluster as being defined by points that are relatively “closer” to one another within the cluster, and “further” from points in another cluster.

One popular set of procedures are hierarchical techniques; whereby either successive fusion or successive divisions are made to create the clusters. The actual specification of the clusters for a hierarchical approach usually requires subjective interpretation of the meaningfulness of the clusters at various levels of the hierarchy. Some of the names that describe these techniques are nearest-neighbor, furthest neighbor, and linkage methods.

A second general set of procedures are called partitioning procedures, where the partitioning is based on an optimization criterion.

One such criterion is based on the generalized variances within and between covariance matrices.

A key feature of most clustering procedures is that there are thresholds that define clusters. At the lowest level each item is its own cluster; at the uppermost level, all the points are in a single cluster. The challenge is to find an intermediate level that permits a reasonable interpretation of its clusters.

Hartigan (1975) provides a general discussion of clustering.
6.5 Classification and discrimination

Classification or discriminant analysis is another classically important problem in which multivariate data is traditionally reduced in complexity. Suppose that there are $k$ populations of individuals, where for each individual we observe $p$ variables $(X_1, \ldots, X_p)$. Based upon historical or "training" data where we observe a sample of $n_i$ individuals from population $i$, $i = 1, \ldots, k$, we want to classify future individuals into one of the $k$ populations based upon observing the $p$ variables for that new individual. Although fairly general procedures have been developed, it has been traditional to assume that for population $i$, $X \sim \mathcal{N}(\mu_i, \Sigma)$, $i = 1, \ldots, k$. Then $\mu_1, \ldots, \mu_k$, $\Sigma$ are estimated from the historical or training data; denote these estimators as $\hat{\mu}_1, \ldots, \hat{\mu}_k, \hat{\Sigma}$. For a new individual, with data $Z$ we want to optimally classify that individual into exactly one of the $k$ populations. In addition, based upon a priori knowledge, the probability that a randomly chosen new individual is from population $i$ is $\pi_i$, $i = 1, \ldots, k$ ($\pi_1 + \cdots + \pi_k = 1$). To do so, we compute a linear function for each population $i$: $S_i = \hat{\mu}_i (\hat{\Sigma}^{-1} Z' - (1/2) \hat{\mu}_i (\hat{\Sigma}^{-1}) \hat{\mu}_i') + \log \pi_i$ and classify $Z$ to the population for which $S_i$ is the largest.

In the case $k = 2$, the linear function $(\mu_1 - \mu_2) \Sigma^{-1} Z'$ or its sample counterpart, $(\hat{\mu}_1 - \hat{\mu}_2) \hat{\Sigma}^{-1} Z$, is called Fisher's linear discriminant function.

Classification can be generalized to incorporate relative costs of misclassification, whereas our preceding development assumed equal costs of misclassification. To handle settings where we do not assume underlying normal populations, there is the technique of logistic discriminant analysis, which again results in a linear discriminant function for two populations. More recent techniques allow nonlinear discriminant functions and non-parametric discriminant functions having a monotonicity property. The classification and regression-tree (CART) approach yields a branching decision tree for classification, see Breiman, Friedman, Olshen, Stone (1984). Decision tree analysis have found many uses in application, e.g., for medical diagnostics. Other references relating to discrimination are Lachenbruch (1995), Goldstein and Dillon (1978).
6.6 Multidimensional scaling

From a geographical map, one can determine the distances between every city and every other city. The technique, called multidimensional scaling (MDS), is concerned with the

inverse problem: Given a matrix of distances, how can we draw a map?
The distances are usually in \( p \)-dimensional space, and the map is in \( q < p \) dimensional space, so that MDS can be thought of as a reduction technique.

Further, \( q = 2 \) is the most natural dimension for a map.

However, we know from the many projection procedures for drawing a map of the globe (\( p = 3 \)) in \( q = 2 \) dimensions that the resulting maps differ in how one perceives distances.

Two fundamental types of MDS are *metric* and *nonmetric* MDS. Metric MDS assumes that the measures are quantitative with interval properties, so that distances, proximities, or similarities can be computed.

Nonmetric MDS assumes that the measures are qualitative, so that rank orders are used in proximity comparisons.

In each case, different programs or algorithms use different definitions for optimal scaling.

For more details see Kruskal and Wish (1978).

6.7 Data Mining

Data mining refers to a set of approaches and techniques that permit “nuggets” of valuable information to be extracted from vast and loosely structured multiple data bases. For example, a consumer products manufacturer might use data mining to better understand the relationship of a specific product’s sales to promotional strategies, selling store’s characteristics, and regional demographics. Techniques from a variety of different disciplines are used in data mining. For instance, computer science and information science
provide methods for handling the problems inherent in focusing and merging the requisite data from multiple and differently-structured data bases. Engineering and economics can provide methods for pattern recognition and predictive modeling. Multivariate statistical techniques, in particular, clearly play a major role in data mining.

Multivariate notions developed to study relationships provide approaches to identify variables or sets of variables that are possibly connected. Regression techniques are useful for prediction. Classification and discrimination methods provide a tool to identify functions of the data that discriminate among categorizations of an individual that might be of interest. Another very useful technique for data mining is cluster analysis that groups experimental units which respond similarly. Structural methods such as principal components, factor analysis, and path analysis are methodologies that can allow simplification of the data structure into fewer important variables. Multivariate graphical methods can be employed to both explore databases and then as a means for presentation of the data mining results.

A good reference to broad issues in data mining is given by Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy (1996).
References


