

- Goodman L A 1986 Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* **54**: 243–70
- Goodman L A, Kruskal W H 1979 *Measures of Association for Cross Classifications*. Springer, New York
- Heinen T 1993 *Discrete Latent Variable Models*. Tilburg University Press, The Netherlands
- Hosmer D W, Lemeshow S 2000 *Applied Logistic Regression*, 2nd edn., Wiley, New York
- Lauritzen, S L 1996 *Graphical Models*. Oxford University Press, Oxford
- Long J S 1997 *Regression Model for Categorical and Limited Dependent Variables*. Sage, Thousand Oaks, CA
- McCullagh P 1980 Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42**: 109–42
- Rabe-Hesketh S, Everitt B 2000 *A Handbook of Statistical Analyses Using Stata*, 2nd edn. CRC Press, London
- Stokes M E, Davis C E, Koch G G 1995 *Categorical Data Analysis Using the SAS System*. SAS, Cary, NC
- Venables W N, Ripley B D 1999 *Modern Applied Statistics with S-Plus*, 3rd edn. Springer, New York

A. Agresti

Multivariate Analysis: Overview

1. Introduction

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.

Although there are many practical applications for each of the methods discussed in this overview, we cite some applications for the classification and discrimination methods in Sect. 6.5. The goal is to distinguish between two populations, or to classify a new observation in one of the populations. Examples are (a) solvent and insolvent companies based on several financial measures; (b) nonulcer dyspeptics versus normal individuals based on measures of anxiety, dependence, guilt, and perfectionism; (c) Alaskan vs.

Canadian salmon based on measures of the diameters of rings. For other such applications, see Johnson and Wichern (1999).

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.

2. Normal Theory

2.1 Distribution Theory

2.1.1 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 $\mathbf{X} = (X_1, \dots, X_p)$ be a p -dimensional vector of the multiple random variables measured on an experimental unit. The probability density, $f(\mathbf{x})$, of \mathbf{X} is multivariate normal, denoted $N(\boldsymbol{\mu}, \Sigma)$, if

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

for $-\infty < x_i < \infty$, $i = 1, \dots, p$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)$ is the p -dimensional population mean vector ($-\infty < \mu_i < \infty$, $i = 1, \dots, p$) and Σ is the $p \times p$ population (positive definite) variance-covariance matrix. When Σ is singular, the probability density function $f(\mathbf{x})$ does not exist as noted, but there are other meaningful ways to define the multivariate normal distribution.

The contours of constancy of $f(\mathbf{x})$ are ellipsoids centered at $\boldsymbol{\mu}$ with axes determined by the eigenvectors of Σ . Any subset of X_1, \dots, 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.

2.1.2 Sample statistics and their properties. Suppose that a random sample $\mathbf{X}_1, \dots, \mathbf{X}_N$, of size N , is observed from a multivariate normal distribution. The natural estimators of $\boldsymbol{\mu}$ and Σ are, respectively, the sample mean vector $\bar{\mathbf{X}}$ and the sample variance-covariance matrix $S = \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' / (N-1)$.

These estimators possess many of the properties of the univariate sample mean and variance. They are independent and unbiased; furthermore, $\bar{\mathbf{X}}$ and $[(N-1)/N]S$ are maximum likelihood estimators. The

sampling distribution of $\bar{\mathbf{X}}$ is $N(\boldsymbol{\mu}, \Sigma/N)$, and the sampling distribution of $(N-1)S$ is called the Wishart distribution with parameter Σ and $N-1$ degrees of freedom, as obtained by John Wishart in 1928.

For the univariate normal distribution, when the variance σ^2 is known, the mean \bar{x} is an optimal estimator of the population mean μ in any reasonable sense. It was shown in 1956 by Charles Stein that when $p > 2$, and the underlying multivariate distribution is $N(\boldsymbol{\mu}, I)$, one can improve upon $\bar{\mathbf{X}}$ as an estimator of $\boldsymbol{\mu}$ 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 $\boldsymbol{\mu}$.

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

2.1.3 Characteristic roots. When the sample covariance matrix S has a Wishart distribution, 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 population covariance matrix was developed by Alan James in 1956 and requires expansions in terms of zonal polynomials. For details of the noncentral distribution of the characteristic roots and zonal polynomials see Mathai et al. (1995).

2.2 Inference

Inference concerning $\boldsymbol{\mu}$ when Σ is known is based, in part, upon the Mahalanobis distance $N(\bar{\mathbf{X}} - \boldsymbol{\mu})' \Sigma^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ which has a χ_N^2 distribution when $\mathbf{X}_1, \dots, \mathbf{X}_N$ is a random sample from $N(\boldsymbol{\mu}, \Sigma)$. When Σ is not known, inference about $\boldsymbol{\mu}$ utilizes the Mahalanobis distance with Σ replaced by its estimator S . The distribution of the quantity $N(\bar{\mathbf{X}} - \boldsymbol{\mu})' S^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ 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 α , the null hypothesis $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against the alternative $H_A: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ based upon multivariate normal data with unknown Σ , one rejects H_0 if $N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' S^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$ exceeds $(N-1)p/(N-p)F_{p, N-p}^\alpha$, where $F_{p, N-p}^\alpha$ denotes the upper α -critical value of $F_{p, N-p}$. Elliptical confidence intervals for $\boldsymbol{\mu}$ can be constructed as $\{\boldsymbol{\mu}: N(\bar{\mathbf{X}} - \boldsymbol{\mu})' S^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq ((N-1)p/(N-p)) \times F_{p, N-p}^\alpha\}$. Useful rectangular confidence intervals for $\boldsymbol{\mu}$ 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. A particular class of techniques that is useful in this case are those of order-restricted inference.

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 -statistic 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.

Source	Sum of squares	Degrees of freedom	F -statistic
Hypothesis	SS_H	d_H	$F = (SS_H/d_H)/(SS_E/d_E)$
Error	SS_E	d_E	

The null hypothesis reference distribution for the F statistic 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 $SS_E \sim \sigma^2 \chi_{d_E}^2$ and under H_0 , $SS_H \sim \sigma^2 \chi_{d_H}^2$. 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 best-known statistics are the Wilks lambda, Hotelling-Lawley trace, Pillai trace, and Roy maximum, and are provided in standard MANOVA statistical packages.

A classical archeological example from 1935 consists of four series of Egyptian skulls ranging from the predynastic, sixth to twelfth dynasties, twelfth and thirteenth dynasties, and Ptolemaic dynasties. For each of the multiple skulls in the four series, there were measurements on breadth, length, and height. A MANOVA showed that the dynasties were not homogeneous. Subsequent analyses were carried out using multiple comparisons. The MANOVA of only two dynasty comparisons reduces to the two-sample Hotelling's T^2 .

3. Relationships

3.1 Regression

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

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 Σ , and the expected value of Y is a function of unknown parameters, namely, $EY = X_1BX_2$, 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 $(\mathbf{X}_1, \dots, \mathbf{X}_p)$ is a p -dimensional random vector with covariances σ_{ij} , the population correlation coefficient between X_i and X_j is defined as $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$, with the sample analogue correspondingly defined. The term co-relation was used by Sir Francis Galton in 1888. Karl Pearson introduced the sample correlation coefficient in 1896; alternative names are product moment correlation or Pearson's correlation coefficient. 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 $\sqrt{N-2}r/\sqrt{1-r^2}$ 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 $z = \tanh^{-1}(r)$, which he showed to have an approximate normal distribution with mean $\tanh^{-1}(\rho)$ and variance $(N-3)^{-1}$. Consequently, the z transformation provides confidence intervals for ρ . The definitive analysis of the distribution of r and z is given by Hotelling (1953).

The population multiple correlation coefficient arises when X_2, \dots, 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, \dots, 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, \dots, 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 the relationship between two sets of variables and are described more fully in Sect. 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. Popular and useful population measures for continuous random variables include Kendall's $\tau (= 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x, y)f(x, y)dx dy - 1)$ and Spearman's $\rho (= 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x, y) - F_x(x)F_y(y))f_x(x)f_y(y)dx dy)$, 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 τ 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 τ and Spearman's rho are given in Agresti (1984). Other more theoretical approaches to measuring dependence are based upon maximizing the correlation, $\rho(a(X), b(Y))$, over certain classes of functions $a(x)$ and $b(y)$. A classical reference is

Goodman and Kruskal (1979); see also *Multivariate Analysis: Discrete Variables (Overview)*.

A related notion is the canonical decomposition of a bivariate distribution into weighted sums of independent-like components (see Lancaster 1969).

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 to generate multivariate distributions is to develop 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 spherical distributions. Here $f(x_1, \dots, x_p) = g(|x_1|, \dots, |x_p|)$ or $f(x_1, \dots, x_p) = g(x_1^2 + \dots + x_p^2)$ 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 are 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 models with non-negative random variables. The introduction of a dependence structure can be ac-

complished by various notions of mixing. For example, if U, V, W are independent random variables; then the ‘mixtures’ (a) $X_1 = U + V, X_2 = U + W$, (b) $X_1 = \min(U, V), X_2 = \min(U, W)$, (c) $X_1 = U/V, X_2 = U/W$ generate dependent bivariate random variables. In particular, a bivariate binomial distribution, a bivariate exponential distribution, and a bivariate Pareto or F distribution can be so generated. An inherent problem with multivariate versions obtained from mixture methods, as well as other generating techniques, is that the number of parameters so introduced can be as large as 2^p , thereby necessitating large sample sizes for estimation. For a comprehensive treatment of bivariate distributions see Hutchinson and Lai (1990).

4.3 Copulas

Another approach to providing non-normal multivariate 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,

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

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

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

Based upon copulas like these, different multivariate distributions can be generated. For instance, the standardized bivariate normal distribution function with correlation ρ 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.

Every 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 known as, 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. Copulas can be readily extended to higher dimensions. However, bounds become more complex. For more details see Rüschendorf et al. (1996).

5. Multivariate Probability Computational Techniques

5.1 Multivariate Bounds for Probabilities

Inequalities or bounds for multivariate probabilities 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. 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. These bounds have many extensions; see Hochberg and Tamhane (1987), Galambos and Simonelli (1996), and *Hypothesis Tests, Multiplicity of*.

A second set of inequalities are called Chebyshev inequalities. For bivariate random variables X_1, X_2 with $EX_i = \mu_i$, $\text{Var}(X_i) = \sigma_i^2$ and correlation

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

thereby extending the classical Bienaymé–Chebyshev univariate inequality.

Another class of inequalities are multiplicative inequalities that hold for certain classes of random variables, namely,

$$P\{X_1 \in A_1, \dots, X_m \in A_m\} \geq \prod_{i=1}^m P\{X_i \in A_i\}$$

When X_1, \dots, X_m are independent normal variates with respective means μ_1, \dots, μ_m , and common variance σ^2 , then

$$P\{\cap |X_i - \mu_i| \geq k_i s\} \geq \prod_{i=1}^m P\{|X_i - \mu_i| \geq k_i s\}$$

where s^2 is an independent estimate of σ^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, \dots, 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 principles; here $P\{\mathbf{x} \in A\} = \int \cdots \int_{\mathbf{x} \in A} f(\mathbf{x}) d\mathbf{x}$ is approximated by $\sum_{i \in \{x_i\}} w_i f(\mathbf{x}_i)$, 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 \mathbf{x}_i , $i = 1, \dots, N$ according to $f(\mathbf{x})$ and estimating $P\{A\}$ by $N^{-1} \sum_{i=1}^N I(\mathbf{x}_i \in A)$, where $I(\mathbf{x}_i \in A) = 1$, if $\mathbf{x}_i \in A$, and $= 0$, otherwise. Some methods for computing multivariate normal probabilities first transform the random variables, and rewrite $P\{A\} = \int \cdots \int_{\|\mathbf{z}\| \leq 1} F(\mathbf{z}) d\mathbf{z}$, where $F(\mathbf{z})$ is a function defined by Σ and the region A . Then $P\{A\}$ is estimated by $N^{-1} \sum_{i=1}^N F(\mathbf{z}_i)$ where the \mathbf{z}_i are randomly chosen from the surface of a sphere in p -dimensional space.

Monte Carlo importance sampling techniques assume a convenient density $g(\cdot)$ whose support includes the support of the density $f(\cdot)$, and $g(\cdot)$ is chosen to expedite convergence. Then points $\mathbf{y}_1, \dots, \mathbf{y}_N$ are random samples according to $g(\mathbf{y})$ and $P\{A\}$ is estimated by $N^{-1} \sum_{i=1}^N I(\mathbf{y}_i \in A) [f(\mathbf{y}_i)/g(\mathbf{y}_i)]$.

More recently, Markov chain Monte Carlo methods have been developed and have proven to be particularly useful for high dimensional random vectors. The common aspects of these types of techniques is the definition of a Markov chain $\mathbf{x}_1, \mathbf{x}_2, \dots$ which converges to a stationary Markov chain where \mathbf{x}_i has marginal distribution $f(\mathbf{x})$. Again, $P\{A\}$ is approximated by $(N - m + 1)^{-1} \sum_{i=m}^N I(\mathbf{x}_i \in A)$, where $\mathbf{x}_m, \dots, \mathbf{x}_N$ is a suitably chosen sequence of elements generated by the Markov chain. For much more complete details, see *Monte Carlo Methods and Bayesian Computation: Importance Sampling; Markov Chain Monte Carlo Methods; Monte Carlo Methods and Bayesian Computation: Overview*.

6. Structures and Patterns

6.1 Principal Components

A correlation matrix represents the interdependencies among 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 (see *Hotelling, Harold (1895–1973)*), methodology that transforms the original correlated standardized variables z_1, \dots, z_p into new variables z_1^*, \dots, z_p^* (called principal components), 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 τ_i^2 ; then the total variance $p = \sum \tau_i^2$ remains fixed for the two systems of variables. By ordering the variances $\tau_{(1)}^2 \geq \tau_{(2)}^2 \geq \dots \geq \tau_{(p)}^2$, and examining the cumulative proportions $\tau_{(1)}^2/p, (\tau_{(1)}^2 + \tau_{(2)}^2)/p, \dots, (\tau_{(1)}^2 + \dots + \tau_{(p)}^2)/p = 1$, 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 provides 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.

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 multiple measurements on two siblings, then canonical analysis provides another exploratory data reduction procedure. If the two sets are labeled X_1, \dots, X_p and Y_1, \dots, 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_i, V_i) , $i = 1, \dots, p$, of the X_i ’s and Y_j ’s, respectively, with the property that all the variances are 1, $\rho(U_i, V_i) = \rho_i$ and all other correlations are zero. The new variables are called canonical variables, and the correlations ρ_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 much of the structure between the two sets of variables.

6.3 Latent Structure and Causal Models

Latent structure models refers to 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. See also *Factor Analysis and Latent Structure: Overview*.

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 multiway 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 details concerning graphical models see Lauritzen (1996) and *Graphical Models: Overview*.

6.3.1 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.

Factor analysis has been used in two data analytic contexts: in a confirmatory manner designed to confirm or negate the hypothesized structure, or to try to discover a structure, in which case the analysis is called exploratory. See also *Latent Structure and Casual Variables; Factor Analysis and Latent Structure, Confirmatory; Factor Analysis and Latent Structure: IRT and Rasch Models; Factor Analysis and Latent Structure: Overview*.

6.3.2 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 sat-

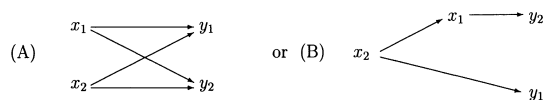


Figure 1
Examples of path structures

isfaction, 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.

In Fig. 1A, variation in x_1 and x_2 affect both outcomes y_1 and y_2 , whereas in Fig. 1B, verbal intelligence affects performance directly, and affects job satisfaction through achievement motivation. Additionally, see *Latent Structure and Casual Variables*.

6.3.3 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 $\mathbf{x} = (x_1, \dots, x_q)$ and $\mathbf{y} = (y_1, \dots, y_p)$ of observable variables, $\mathbf{y} = \boldsymbol{\eta}\boldsymbol{\Lambda}_1 + \boldsymbol{\varepsilon}$, $\mathbf{x} = \boldsymbol{\xi}\boldsymbol{\Lambda}_2 + \boldsymbol{\delta}$, connected by $\boldsymbol{\eta}\boldsymbol{\Delta} = \boldsymbol{\xi}\boldsymbol{\Gamma} + \boldsymbol{\zeta}$ where $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$ are, respectively, $m \times p$ and $q \times n$ matrices of regression coefficients; $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are vectors of unobservable dependent variables, $\boldsymbol{\Delta}$ is an $m \times m$ regression matrix of the effects of endogenous on endogenous variables, $\boldsymbol{\Gamma}$ is an $m \times n$ regression matrix of the effects of exogenous on endogenous variables, and $\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$ are residual vectors. For more details see *Latent Structure and Casual Variables*.

6.4 Clustering

Often, subsets of individuals are grouped by a known characteristic, as for example, by sex, by race, by socioeconomic 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 individual (objects). For each pair of individuals, we form a ‘distance’ measure, where ‘distance’ is flexibly defined. Thus we create an $n \times n$ matrix of ‘distances’ for which 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. See *Statistical Clustering* for more details.

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, \dots, X_p) . Based upon historical or ‘training’ data where we observe a sample of n_i individuals from population i , $i = 1, \dots, k$, we want to develop a method 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 , $\mathbf{X} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \dots, k$. Then $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ are estimated from the historical or training data. For a new individual, with data \mathbf{Z} we want to optimally classify that individual into exactly one of the k populations.

The case $k = 2$ yields a linear classifier involving \mathbf{Z} called Fisher’s linear discriminant function, and the case $k > 2$ involves k linear functions of \mathbf{Z} .

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 nonparametric discriminant functions having a monotonicity property. The classification and regression-tree (CART) approach yields a branching decision tree for classification; see Breiman et al. (1984). Decision tree analyses have found many uses in applications, e.g., for medical diagnostics. For further information on discrimination see *Multivariate Analysis: Classification and Discrimination*.

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 dimension 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. For more details see *Scaling: Multidimensional*.

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 reference to broad issues in data mining is given by Fayyad et al. (1996). Also see *Exploratory Data Analysis: Multivariate Approaches (Nonparametric Regression)*.

7. Further Readings

This article and related entries provide only a brief overview of a vast field. For further general readings see Anderson (1984), Dillon and Goldstein (1989), Johnson and Wichern (1999), Morrison (1990), Muirhead (1982); for detailed developments see the *Journal of Multivariate Analysis*.

Bibliography

- Agresti A 1984 *Analysis of Ordinal Categorical Data*. Wiley, New York
 Anderson T W 1984 *An Introduction to Multivariate Statistical Analysis*, 2nd edn. Wiley, New York
 Breiman L, Friedman J H, Olshen R A, Stone C J 1984 *Classification and Regression Trees*. Wadsworth, Belmont, CA

- Dharmadhikari S W, Joag-dev K 1988 *Unimodality, Convexity, and Applications*. Academic Press, Boston
 Dillon W R, Goldstein M 1989 *Multivariate Analysis: Methods and Applications*. Wiley, New York
 Fang K-T, Anderson T W (eds.) 1990 *Statistical Inference in Elliptically Contoured and Related Distributions*. Allerton Press, New York
 Fayyad U M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds.) 1996 *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence/MIT Press, Cambridge, MA
 Galambos J, Simonelli E 1996 *Bonferroni-type Inequalities with Applications*. Springer-Verlag, New York
 Goodman L, Kruskal W 1979 *Measures of Association for Cross Classifications*. Springer-Verlag, New York
 Hochberg Y, Tamhane A C 1987 *Multiple Comparison Procedures*. Wiley, New York
 Hotelling H 1953 New light on the correlation coefficient and its transform. *Journal of the Royal Statistical Society, Series B* **15**: 193–225
 Hutchinson T P, Lai C D 1990 *Continuous Bivariate Distributions, Emphasizing Applications*. Rumsby Scientific Publishing, Adelaide, South Australia
 Joe H 1997 *Multivariate Models and Dependence Concepts*. Chapman & Hall, London
 Johnson R A, Wichern D W 1999 *Applied Multivariate Statistical Analysis*, 4th edn. Prentice-Hall, New York
 Lancaster H O 1969 *The Chi-squared Distribution*. Wiley, New York
 Lauritzen S L 1996 *Graphical Models*. Oxford University Press, Oxford, UK
 Mathai A M, Provost S B, Hayakawa T 1995 *Bilinear Forms and Zonal Polynomials*. Springer Verlag, New York
 Morrison D F 1990 *Multivariate Statistical Methods*, 3rd edn. McGraw-Hill, New York
 Muirhead R J 1982 *Aspects of Multivariate Statistical Theory*. Wiley, New York
 Nelsen R B 1999 *An Introduction to Copulas*. Springer, New York
 Rüschendorf L, Schweizer B, Taylor M D (eds.) 1996 *Distributions with Fixed Marginals and Related Topics*. Lecture Notes – Monograph Series, Vol. 28, Institute of Mathematical Statistics, Hayward, CA
 Tong Y L 1980 *Probability Inequalities in Multivariate Distributions*. Academic Press, New York
 Tong Y L 1990 *The Multivariate Normal Distribution*. Springer-Verlag, New York

I. Olkin and A. R. Sampson

Copyright © 2001 Elsevier Science Ltd.

All rights reserved.

Mumford, Lewis (1895–1990)

Lewis Mumford was one of the seminal cultural critics to emerge from the United States during the twentieth century. Although known chiefly for his work in architecture and urbanism, he made significant contributions to art and literary criticism, biography and autobiography, the history of technology, sociology, and political science. Largely self-taught, Mumford