Panel on Discriminant Analysis, Classification, and Clustering 1989 Discriminant analysis and clustering. *Statistical Science* **4**: 34–69
Quinlan J R 1993 *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA
Ripley B D 1996 *Pattern Recognition and Neural Networks*. Cambridge University Press, New York
Vapnik V N 2000 *The Nature of Statistical Learning Theory*, 2nd edn. Springer, New York

G. McLachlan

Multivariate Analysis: Discrete Variables (Correspondence Models)

1. Introduction

Correspondence models extract information on the association between categorical variables. As in correspondence analysis (see *Scaling: Correspondence Analysis*), correspondence models are based on a canonical parametrization of the joint distribution of the variables. Traditional correspondence analysis (Benzécri 1973, Greenacre 1984, Hill 1974, Nishisato 1980, Gifi 1990) is concerned with methods used for description and detection of patterns of association between variables via graphical displays, and does not consider the role of random sampling in the generation of the observed data from a population with an associated probability model. In contrast, correspondence models are explicitly concerned with inferential methods for the study of probability models which are fitted to observed data under the assumption that the data have been obtained by random sampling (Goodman 1985, 1986, Gilula and Haberman 1986, 1988). The parameterizations used in correspondence models may still be used to construct descriptive measures and graphs; however, unlike in traditional correspondence analysis, correspondence models involve formal statistical methods such as tests of goodness of fit and approximate confidence intervals for parameters. Development of correspondence models is most straightforward in the case in which the relationship of two categorical random variables is studied (Sect. 2). When more than two variables are involved, correspondence models may involve a division of the categorical variables into two groups (Sect. 3). Correspondence models based on multiple correspondence analysis are much more difficult to construct (Sect. 4).

2. Bivariate Correspondence Models

Let X and Y be two categorical random variables with respective ranges consisting of I and J categories indexed from 1 to I, and from 1 to J, respectively. The joint probability distribution of X and Y may be characterized by the I by J matrix p^{XY} of joint probabilities $p_{ij}^{XY} = P(X = i, Y = j)$, $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$. The marginal probability distributions of X and Y may be characterized by the marginal probabilities $p_i^X = P(X = i)$, $1 \leqslant i \leqslant I$, and $p_j^Y = P(Y = j)$, $1 \leqslant j \leqslant J$. For simplicity, assume that $p_i^X > 0$ for $1 \leqslant i \leqslant I$ and $p_j^Y > 0$ for $1 \leqslant j \leqslant J$. The entire family of correspondence models for bivariate distributions is based on the canonical decomposition (Hirschfeld 1935, Fisher 1940)

$$p_{ij}^{XY} = p_i^X p_j^Y \left(1 + \sum_{m=1}^{M} \rho_{im} \mu_{im} \nu_{jm} \right) \tag{1}$$

where $M = \min(I-1, J-1)$,
$1 \geqslant \rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant \rho_M \geqslant 0$

$$\sum_{i=1}^{I} p_i^X \mu_{im} = \sum_{j=1}^{J} p_j^Y \nu_{jm} = 0, 1 \leqslant m \leqslant M \tag{2}$$

and

$$\sum_{i=1}^{I} p_i^X \mu_{im}\mu_{im'} = \sum_{j=1}^{J} p_j^Y \nu_{jm}\nu_{jm'} = \begin{cases} 1, 1 \leqslant m = m' \leqslant M \\ 0, 1 \leqslant m \leqslant m' \leqslant M \end{cases} \tag{3}$$

In (1), (2), and (3), for each integer m from 1 to M, the parameter ρ_m is the mth canonical correlation, the score μ_{im} is the mth canonical score of category i of variable X, $1 \leqslant i \leqslant I$, and the score ν_{jm} is the mth canonical score of category j of variable Y. Let the mth canonical variable μ_{Xm} associated with X be the random variable with value μ_{im} for $X = i$, and let the mth canonical variable μ_{Ym} associated with Y denote the random variable with value ν_{jm} for $Y = j$. Then μ_{Xm} and ν_{Ym} both have mean 0 and variance 1, and ρ_m is the correlation of the canonical variables μ_{Xm} and ν_{Ym}. In addition, if $1 \leqslant m < m' \leqslant M$, then the canonical variables μ_{Xm} and $\mu_{Xm'}$ of X are uncorrelated, and the canonical variables ν_{Ym} and $\nu_{Ym'}$ of Y are uncorrelated.

For some M^* between 1 and M, graphical display of category i of X may be based on the observed scores μ_{im} for $1 \leqslant m \leqslant M^*$ and graphical display of category j of Y may be based on the observed scores ν_{jm} for $1 \leqslant m \leqslant M^*$. An alternative approach uses graphical displays for row category i based on the correspondence scores $\rho_m \mu_{im}$ for $1 \leqslant m \leqslant M^*$ and graphical displays for column category j based on the correspondence scores $\rho_m \nu_{jm}$ for $1 \leqslant m \leqslant M^*$.

Formal inference for canonical models has generally been based on maximum likelihood. Computation of maximum-likelihood estimates has been attempted by use of a cyclic algorithm (Goodman 1985) and by use of a scoring algorithm for parameters subject to constraints (Gilula and Haberman 1986). Goodness of fit has typically been examined by use of likelihood-ratio and Pearson chi-square statistics. In typical cases, a chi-square approximation can be applied when the model is true. If the model only assumes that $\rho_m = 0$

Table 1
Cross-classification of origin by principal worry

	Principal worry[a]							
Origin	ENR	SAB	MIL	POL	ECO	OTH	MTO	PER
A/A	61	70	97	32	4	81	20	104
E/A	104	117	218	118	11	128	42	48
fA/A	8	9	12	6	1	14	2	14
fE/A	22	24	28	28	2	52	6	16
fI/I	5	7	14	7	1	12	0	9

Source: Guttman (1971).
[a]Columns are: ENR, enlisted relative; SAB, sabotage; MIL, military situation; POL, political situation; ECO, economic situation; OTH, other; MTO, more than one worry; and PER, personal economics. Rows are: A/A, Africa or Asia; E/A, Europe or America; fA/A, father from Africa or Asia; fE/A, father from Europe or America; and fI/I, father and self from Israel.

Table 2
Estimated canonical parameters for model with two pairs of scores

Principal worry	Score 1	Score 2
ENR	−0.0573	−0.2380
SAB	−0.0096	−0.2880
MIL	−0.5121	−1.0690
POL	−1.2270	0.6685
ECO	−0.6866	0.0811
OTH	0.0692	1.7794
MTO	−0.5203	−1.1601
PER	2.4428	−0.2684
Origin	**Score 1**	**Score 2**
A/A	1.3472	−0.4105
E/A	−0.8663	−0.4678
fA/A	1.1632	0.7179
fE/A	−0.2830	2.6321
fI/I	0.4120	0.8064

for $m > M^*$, then the chi-square approximations have $(I - M^* - 1)(J - M^* - 1)$ degrees of freedom.

As a statistical model, (1) is a *saturated* model, for no restrictions are imposed on the joint probabilities p_{ij}^{XY}. Goodman (1985, 1986) and Gilula and Haberman (1986) consider models in which the canonical correlations ρ_m are assumed 0 for $m > M^*$ for some M^* between 0 and M. The case of $M^* = 0$ leads to the conventional model of independence of the variables X and Y. The case of $M^* = M$ is the saturated model in which no restrictions are imposed. The most common nontrivial case has $M^* = 1$. Models are also considered in which linear restrictions are imposed on the scores μ_{im} and v_{jm} for $m \leqslant M^*$.

As an example, consider the data in Table 1 (Guttman 1971) on the origin of Israeli adults versus their principal worry. Several cells in the table have small values, so some care must be exercised in interpretation of customary large-sample approximation. Nonetheless, it is worth noting that the model with a single nonzero canonical correlation ($\rho_m = 0$ for $m > 1$) yields a likelihood-ratio chi-square of 29.35

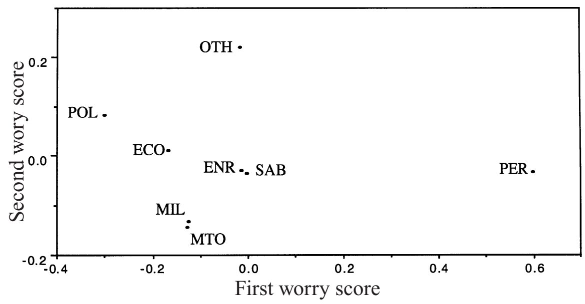

Figure 1
Correspondence analysis scores for worry

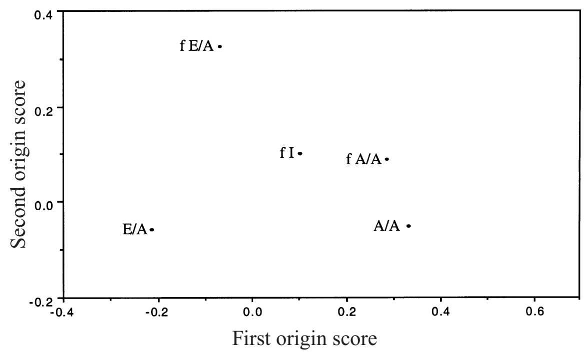

Figure 2
Correspondence analysis scores for origin

on 18 degrees of freedom, so that the fit is relatively poor. On the other hand, the model with two nonzero canonical correlations ($\rho_m = 0$ for $m > 2$) yields a likelihood-ratio chi-square of only 6.49 on 10 degrees of freedom, so that the fit is quite satisfactory. The estimated canonical correlations are 0.245 and 0.124. These results suggest a modest relationship between the variables under study. The maximum-likelihood estimates of the canonical scores are provided in Table 2. Fig. 1 provides correspondence scores for the column variable, and Fig. 2 provides correspondence scores for the row variable. Results suggest a contrast between personal worries and other concerns, especially political ones, with subjects of Asian or African background more likely to express personal worries than subjects of European or American background. The second dimension of background appears to involve a contrast between subjects born in Israel with fathers of European or American background and other subjects. Subjects born in Israel with fathers from Europe or America are more likely to cite more than one worry than are other subjects.

3. *Two Groups of Categorical Variables*

Extension of correspondence models to apply to more than two categorical variables is a nontrivial matter. The simplest approach is appropriate if the variables can be divided into two groups (Gilula and Haberman 1988). In this case, each group of variables can be regarded as a categorical variable with many categories, and correspondence analysis methods for the bivariate case can be applied. For example, in addition to the categorical random variables X and Y of Sect. 2 let an additional categorical random variable Z be defined with values from 1 to $K \geqslant 2$. As an example, Gilula and Haberman (1988) consider an example with X equal to education level, Y equal to religion, and Z equal to attitude toward legal nontherapeutic abortions. Each variable in this example involves a classification into three categories, so that I, J, and K are 3. In general, let p_{ijk}^{XYZ} be the joint probability that $X = i$, $Y = j$, and $Z = k$. Then

$$p_{ijk}^{XYZ} = p_{ij}^{XY} p_k^Z \left(1 + \sum_{m=1}^{M} \rho'_m \mu'_{ijm} v'_{km} \right), \tag{4}$$

where M is the minimum of $IJ-1$ and $K-1$, the ρ'_m are nonnegative and nonincreasing in m,

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}^{XY} \mu'_{ijm} = \sum_{k=1}^{K} p_k^Z v'_{km} = 0, \quad 1 \leqslant m \leqslant M, \tag{5}$$

and

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}^{XY} \mu'_{ijm} ? \mu'_{ijm'} = \sum_{k=1}^{K} p_k^Z v'_{km} v'_{km'} = \begin{cases} 1, 1 \leqslant m = m' \leqslant M, \\ 0, 1 \leqslant m < m' \leqslant M. \end{cases} \tag{6}$$

Various restrictions may be imposed on the canonical parameters. For example, if Y and Z are conditionally independent given X, then $\rho_m = 0$ for any $m > (-1)$ and μ'_{ijm} depends only on i for $\rho_m > 0$. For other examples, see Gilula and Haberman (1988).

4. *Multiple Correspondence Analysis*

Multiple correspondence analysis, another approach to extension of correspondence analysis to the study of two or more categorical variables, appears in Guttman (1941). Multiple correspondence analysis can be regarded as a special case of correspondence analysis. For illustrative purposes, the case of two and three variables will be considered.

Consider the case of two categorical random variables X and Y defined as in Sect. 2. In a multiple correspondence analysis of X and Y, a canonical analysis is used to predict the joint variable (X, Y) from either X or Y, where X or Y is chosen at random. Formally, (X, Y) is predicted by the pair (U, V), where V is a random variable independent of X and Y such that V assumes the values 1 and 2 with equal probability. The variable U is X if V is 1, and U is Y if V is 2. when $V = 1$, then $U = X$, so that prediction of (X, Y) reduces to prediction of Y by X. When $V = 2$, then $U = Y$, so that prediction of (X, Y) reduces to prediction of X by Y. The canonical analysis in this case is closely related to conventional canonical analysis of the variables X and Y.

For multiple correspondence analysis of the three categorical variables X, Y, and Z of Sect. 3 the joint variable (X, Y, Z) is predicted by random choice of a marginal variable X, Y, or Z. In this case, let V be a random variable independent of X, Y, and Z which assumes value k with probability $\frac{1}{3}$ for $1 \leqslant k \leqslant 3$. Let U be the random variable which is X for $V = 1$, Y for $V = 2$, and Z for $V = 3$. Multiple correspondence analysis considers correspondence analysis of (X, Y, Z) and U. If $V = 1$, then $U = X$ and prediction of (X, Y, Z) reduces to prediction of (Y, Z) by X. If $V = 2$, then $U = Y$ and prediction of (X, Y, Z) reduces to prediction of (X, Z) by Y. If $V = 3$, then $U = Z$ and prediction of (X, Y, Z) reduces to prediction of (X, Y) by Z. The analysis differs from that in Sect. 3 in that it is symmetrical in the three variables. Thus in the example of Gilula and Haberman (1988), the analysis looks at educational level, religion, and abortion attitude symmetrically. This approach may or may not be appropriate, depending on the context. For instance, one may reasonably argue that abortion attitude reflects educational and religious background, but neither educational background nor religious background is appreciably affected by abortion attitude.

5. *Conclusion*

Correspondence models compete in practice with a number of other statistical methods in common use. In their development of canonical models, both Goodman (1985) and Gilula and Haberman (1986) compare canonical models with association models in the case of two variables. Here association models involve a loglinear parameterization similar to the parameterization used in correspondence models. Interpretations involve loglinear interactions rather than the correlations used in correspondence models, but it remains the case that scores are attached to variable categories. As noted in both papers, association models in which scores are fixed rather than estimated are special cases of conventional log-linear models (see *Multivariate Analysis: Discrete Variables (Loglinear Models)*). Goodman (1985) also discusses the relationship of parameters in canonical models to parameters in correspondence models. Gilula and Haberman (1986) and van der Heijden, Gilula and van der Ark (1999) consider the relationship of canonical models to latent-class models (see *Factor Analysis and Latent Structure: Overview*). In the case in which more than two variables are analyzed by division of the variables into two groups, Gilula and Haberman (1988) compare canonical models to association models in cases in which more than two variables are present.

The canonical correlations used in bivariate correspondence analysis may be used to derive formal tests of independence that compete with conventional chi-square statistics and may be employed to ascertain the nature of departures from independence. For details, see Haberman (1981).

A variety of alternative interpretations of multiple correspondence analysis are available. For example, see Tenenhaus and Young (1985) and Gifi (1990).

See also: Factor Analysis and Latent Structure: Overview; Multivariate Analysis: Discrete Variables (Loglinear Models); Multivariate Analysis: Overview; Scaling: Correspondence Analysis

Bibliography

Benzécri J-P 1973 L'Analyse des Données: 2. *L'Analyse de Correspondances*. Dunod, Paris

Fisher R A 1940 The precision of discriminant functions. *Annals of Eugenics* **10**: 422–29

Gifi A 1990 *Non-linear Multivariate Analysis*. Wiley, New York

Gilula Z and Haberman S J 1986 Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*. **81** 780–6

Gilula Z, Haberman S J 1988 The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association* **83**: 760–71

Goodman L A 1985 Analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* **13**: 10–69

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**: 243–309

Greenacre M J 1984 *Theory and Applications of Correspondence Analysis*. Academic Press, London

Guttman L 1941 The quantification of a class of attributes: a theory and method of scale construction. In: Horst P (ed.) *The Prediction of Personal Adjustment*. Social Science Research Council, New York, pp. 319–48

Guttman L 1971 Measurement as structural theory. *Psychometrika* **36**: 329–47

Haberman S J 1981 Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Annals of Statistics* **9**: 1178–86

Hill M O 1974 Correspondence analysis: a neglected multivariate method. *Applied Statistics* **23**: 340–54

Hirschfeld H O 1935 A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society* **31**: 520–4

Nishisato S 1980 *Analysis of Categorical Data: Dual Scaling and Its Applications*. University of Toronto Press

Tenenhaus M, Young F W 1995 An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* **50**: 91–119

Van der Heijden P G M, Gilula Z, van der Ark L A 1999 An extended study into the relationships between correspondence analysis and latent class analysis. In: Sobel M, Becker M (eds.) *Sociological Methodology*. Blackwell, Cambridge pp. 147–86

Z. Gilula, S. J. Haberman, and P. G. M. van der Heijden

Multivariate Analysis: Discrete Variables (Logistic Regression)

Logistic regression is used widely to examine and describe the relationship between a binary response variable (e.g., 'success' or 'failure') and a set of predictor variables. In common with linear regression (e.g., *Linear Hypothesis: Regression (Basics)*), the primary objective of logistic regression is to model the mean of the response variable, given a set of predictor variables. However, what distinguishes logistic regression from linear regression is that the response variable is binary rather than continuous in nature. This has a number of consequences for modeling the mean of the response variable. In this article the main features of logistic regression are described and some aspects of interpretation of logistic regression are illustrated with an example.

1. Introduction to Logistic Regression

In this section we introduce some notation and consider regression models for a binary response variable. Let Y denote a binary response variable, that is, a discrete response variable having only two categories, for convenience often referred to as 'success' or 'failure.' For example, Y might indicate an

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7