Multivariate Analysis: Classification and Discrimination

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

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 (Benzecri 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 J by J matrix \( p_{ij} \) of joint probabilities, \( p_{ij} = P(X = i, Y = j) \), \( 1 \leq i \leq I, 1 \leq j \leq J \). The marginal probability distributions of X and Y may be characterized by the marginal probabilities \( p_{i} = P(X = i) \), \( 1 \leq i \leq I \), and \( p_{j} = P(Y = j) \), \( 1 \leq j \leq J \). For simplicity, assume that \( p_{ij} > 0 \) for \( 1 \leq i \leq I \) and \( p_{ij} > 0 \) for \( 1 \leq j \leq J \). The entire family of correspondence models for bivariate distributions is based on the canonical decomposition (Hirschfeld 1935, Fisher 1940)

\[
p_{ij} = p_{i} p_{j} + \sum_{m=1}^{M} \rho_{m} \mu_{im} \nu_{jm}
\]

where \( M = \min(I - 1, J - 1) \),
\[
1 \geq \rho_{1} \geq \rho_{2} \geq \cdots \geq \rho_{M} \geq 0
\]

and
\[
\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \mu_{im} \nu_{jm} = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \nu_{jm} \mu_{im} = \left\{ \begin{array}{ll}
1, & 1 \leq m = m' \leq M \\
0, & 1 \leq m \leq m' \leq M
\end{array} \right.
\]

In (1), (2), and (3), for each integer m from 1 to M, the parameter \( \rho_{m} \) is the nth canonical correlation, the score \( \mu_{im} \) is the nth canonical score of category i of variable X, \( 1 \leq i \leq I \), and the score \( \nu_{jm} \) is the nth canonical score of category j of variable Y. Let the nth canonical variable \( \mu_{im} \) associated with X be the random variable with value \( \mu_{im} \) for \( X = i \), and let the nth canonical variable \( \nu_{jm} \) associated with Y denote the random variable with value \( \nu_{jm} \) for \( Y = j \). Then \( \mu_{im} \) and \( \nu_{jm} \) both have mean 0 and variance 1, and \( \rho_{m} \) is the correlation of the canonical variables \( \mu_{im} \) and \( \nu_{jm} \). In addition, if \( 1 \leq m < m' \leq M \), then the canonical variables \( \mu_{im} \) and \( \mu_{im'} \) of X are uncorrelated, and the canonical variables \( \nu_{jm} \) and \( \nu_{jm'} \) 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 \( \mu_{im} \) for \( 1 \leq m \leq M^* \) and graphical display of category j of Y may be based on the observed scores \( \nu_{jm} \) for \( 1 \leq m \leq M^* \). An alternative approach uses graphical displays for row category i based on the correspondence scores \( \rho_{m} \mu_{im} \) for \( 1 \leq m \leq M^* \) and graphical displays for column category j based on the correspondense scores \( \rho_{m} \nu_{jm} \) for \( 1 \leq m \leq M^* \).

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

<table>
<thead>
<tr>
<th>Origin</th>
<th>ENR</th>
<th>SAB</th>
<th>MIL</th>
<th>POL</th>
<th>ECO</th>
<th>OTH</th>
<th>MTO</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/A</td>
<td>61</td>
<td>70</td>
<td>97</td>
<td>32</td>
<td>4</td>
<td>81</td>
<td>20</td>
<td>104</td>
</tr>
<tr>
<td>E/A</td>
<td>104</td>
<td>117</td>
<td>218</td>
<td>118</td>
<td>11</td>
<td>128</td>
<td>42</td>
<td>48</td>
</tr>
<tr>
<td>fA/A</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>6</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>fE/A</td>
<td>22</td>
<td>24</td>
<td>28</td>
<td>28</td>
<td>2</td>
<td>52</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>fI/A</td>
<td>5</td>
<td>7</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

Source: Guttman (1971).
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/A, father and self from Israel.

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

<table>
<thead>
<tr>
<th>Principal worry</th>
<th>Score 1</th>
<th>Score 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENR</td>
<td>-0.0573</td>
<td>-0.2380</td>
</tr>
<tr>
<td>SAB</td>
<td>-0.0096</td>
<td>-0.2880</td>
</tr>
<tr>
<td>MIL</td>
<td>-0.5121</td>
<td>-1.0690</td>
</tr>
<tr>
<td>POL</td>
<td>-1.2270</td>
<td>0.6685</td>
</tr>
<tr>
<td>ECO</td>
<td>-0.6866</td>
<td>0.0811</td>
</tr>
<tr>
<td>OTH</td>
<td>0.0692</td>
<td>1.7794</td>
</tr>
<tr>
<td>MTO</td>
<td>-0.5203</td>
<td>-1.1601</td>
</tr>
<tr>
<td>PER</td>
<td>2.4428</td>
<td>-0.2684</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Origin</th>
<th>Score 1</th>
<th>Score 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/A</td>
<td>1.3472</td>
<td>-0.4105</td>
</tr>
<tr>
<td>E/A</td>
<td>-0.8663</td>
<td>-0.4678</td>
</tr>
<tr>
<td>fA/A</td>
<td>1.1632</td>
<td>0.7179</td>
</tr>
<tr>
<td>fE/A</td>
<td>-0.2830</td>
<td>2.6321</td>
</tr>
<tr>
<td>fI/I</td>
<td>0.4120</td>
<td>0.8064</td>
</tr>
</tbody>
</table>

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 \( \rho_\cdot \cdot \cdot \). Goodman (1985, 1986) and Gilula and Haberman (1986) consider models in which the canonical correlations \( \rho_\cdot \cdot \cdot \) 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 \( \mu_{\cdot m} \) and \( \nu_{\cdot m} \) for \( m \leq M^* \).

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 (\( \rho_\cdot \cdot \cdot = 0 \) for \( m > 1 \)) yields a likelihood-ratio chi-square of 29.35 on 18 degrees of freedom, so that the fit is relatively poor. On the other hand, the model with two nonzero canonical correlations (\( \rho_\cdot \cdot \cdot = 0 \) 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.
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 personal worries, especially political ones, with subjects of Asian or African background more likely to express personal worries.

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 $K \geq 2$. 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 $p_{ijk}^{X,Y,Z}$ be the joint probability that $X = i$, $Y = j$, and $Z = k$. Then

$$p_{ijk}^{X,Y,Z} = p_{ij}^{X,Y} p_k^Z \left( 1 + \sum_{m=1}^M \rho_m \mu_{ijm} v_{km} \right), \quad (4)$$

where $M$ is the minimum of $IJ - 1$ and $K - 1$, the $\rho_m$ are nonnegative and nonincreasing in $m$,

$$\sum_{i=1}^I \sum_{j=1}^J p_{ij}^{X,Y} \mu_{ijm} = p_m^X \sum_{k=1}^K v_{km} = 0, \quad 1 \leq m \leq M, \quad (5)$$

and

$$\sum_{i=1}^I \sum_{j=1}^J p_{ij}^{X,Y} \mu_{ijm} v_{km} = \sum_{k=1}^K p_k^Z v_{km} v_{km'}, \quad 1, 1 \leq m = m' \leq M,$$

$$\quad \{ \begin{array}{ll} 0, & 1 \leq m \leq M, \end{array} \} \quad (6)$$

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 $\mu_{ijm}$ 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 $1 \leq k \leq 3$. 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

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


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