Journal of Multivariate Analysis 132 (2014) 9–24

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The analysis of distance of grouped data with categorical variables: Categorical canonical variate analysis

Niël J. Le Roux [a,*], Sugnet Gardner-Lubbe [b], John C. Gower [c]

[a] Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa
[b] Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa
[c] Department of Mathematics and Statistics, The Open University, Milton Keynes, UK

ARTICLE INFO

Article history:
Received 14 September 2013
Available online 7 August 2014

AMS subject classifications:
62H99

Keywords:
Analysis of distance
Biplot
Canonical variate analysis
Categorical canonical variate analysis
Category level point
Discriminant analysis
Generalised biplot
Nonlinear biplot
Prediction region
Singular value decomposition

ABSTRACT

We use generalised biplots to develop the important special case of (i) when all variables are categorical and (ii) the samples fall into K recognised groups. We term this Categorical Canonical Variate Analysis (CatCVA), because it has similar characteristics to Rao's Canonical Variate Analysis (CVA), especially its visual aspects. It allows centroids of groups to be exhibited in increasing numbers of dimensions, together with information on within-group sample variation. Variables are represented by category-level-points (CLPs) which are a counterpart of numerically calibrated biplot axes for quantitative variables. Mechanisms are provided for relating the samples to their category levels, for giving convex regions to help predict categories, and for adding new samples. Inter-sample distance may be measured by any Euclidean embeddable distance. Computation is minimised by working in the $K - 1$ dimensional space containing the group centroids.

The methodology is illustrated by an example with three groups and 37 samples but the number of samples size is not a serious limitation. The visualisation of group structure is the main focus of this paper; computational efficiency is a bonus.

1. Introduction

In the multidimensional scaling of a dissimilarity matrix, the pairwise differences between n samples are mapped into n points in some small number of dimensions, r, usually two. Many methods of multivariate analysis fall into a framework, which we term Analysis of Distance (AoD), where the dissimilarity matrix has been derived from a data matrix $\mathbf{X} : n \times p$ of continuous or categorical variables. The important word in the last sentence is "derived" which, of course includes the identity transformation. For continuous variables the display can be augmented by p linear axes or, less usually, p nonlinear trajectories. The resulting display is a biplot interpreted by evaluating inner-products, either directly or, we think more conveniently, by calibrating the axes in the conventional way for coordinate axes. In contrast, a categorical variable j, say, has a finite number of category levels L_j, say, that cannot be handled in the same way as continuous variables. Rather, each category level must be displayed as a single point, known as a category level point or CLP and thus, the jth variable is represented by L_j CLPs. All p categorical variables generate a total of $L = L_1 + L_2 + \cdots + L_p$ CLPs. A general methodology for CLPs has been given by Gower [4], Gower and Hand [5] and Gower, Lubbe and Le Roux [9], giving analytical properties and examples. The essential thing to recall is that while axes (whether linear or nonlinear) provide a reference system for quantitative variables, CLPs provide a similar reference system for categorical variables. In this paper the CLPs are derived

* Corresponding author.
E-mail addresses: njlr@sun.ac.za (N.J. Le Roux), Sugnet.Lubbe@uct.ac.za (S. Gardner-Lubbe), John.Gower@open.ac.uk (J.C. Gower).

http://dx.doi.org/10.1016/j.jmva.2014.07.014
0047-259X/

from Generalised Biplot theory (see the above references); for a discussion of alternative ways of determining Category Points see the discussion of Section 5.

In this paper we are concerned with what modifications are required when

(i) all p variables are categorical and
(ii) the n samples in $\mathbf{X}$ fall into K groups.

Grouped sample-structure is a feature of Canonical Variate Analysis (CVA) of continuous variables and indeed is one of the most useful tools of multivariate analysis (see e.g. [10]). CVA is based on Mahalanobis distance but there have been many generalisations based on continuous variables using general definitions of distance between grouped samples. Gower, Le Roux and Lubbe [8] give an up-to-date account of this work in the context of AoD and provide a comprehensive list of citations. A similar procedure for categorical variables is described below which, although based on similar fundamental geometric ideas, has a very different manifestation when expressed in algebraic form. We term this variant of AoD Categorical CVA or, in short, CatCVA. It is to be noted that here we are not concerned with canonical correlation or any of its cognates such as OVERALS (see [2]) but confine our attention to the grouping of samples and not to the grouping of variables. Although some forms of Canonical Correlation Analysis (CCA) share some of the algebra of CVA, the two are statistically very different; Gardner, Gower, and Le Roux [1] highlight the difference in an analysis which finds a synthesis between CCA and CVA. Also, we are primarily concerned with between group differences and have only secondary concerns with within-group variability. In this we are in line with classical CVA and, as with CVA itself, our results are not relevant for a detailed analysis of intra-group differences.

Categorical variables are familiar in correspondence analysis, and especially multiple correspondence analysis, but categorical variables are more rarely encountered with grouped samples. We suspect that this could be because there is a dearth of suitable methodology; we hope that this paper will encourage greater use of multivariate categorical variables in the context of grouped samples.

As a referee pointed out, categorical variables are inherent even in classical CVA, because the groups themselves define categories (e.g. the three *Ocotea* species discussed in Section 4). Indeed, categorical variables label rows, columns, etc. of all multiway tables, whether the bodies of the tables give values of numerical, or in our case, categorical variables. Often the body of the table can be considered as observed/response variables and the labelling information as allocated dependent variables, though the distinction is by no means absolute.

First, we introduce our notation; then in Section 2 we review the representation of categorical variables in the ungrouped case. Finally, in Section 3 we show how the methodology of the grouped case can be developed to accommodate both between and within group information. An example is discussed in Section 4.

Notation

$\mathbf{X} \equiv \{x_{ij}\}$	is an $n \times p$ data-matrix for n cases (samples) and p categorical variables.
$\mathbf{Y} \equiv \{y_{ij}\}$	is an $n \times m$ matrix of coordinates.
$\mathbf{G} \equiv \{g_{ik}\}$	is an $n \times K$ matrix indicating membership of K groups. Element $g_{ik} = 1$ when the ith case is in group k, else $g_{ik} = 0$.
$\mathbf{g}_k$	is the kth column of $\mathbf{G}$.
$\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_p]$	is an $n \times L$ indicator matrix associated with the p categorical variables where $\mathbf{C}_j : n \times L_j$ denotes the indicator matrix associated with the jth variable having L_j different category levels and $L = L_1 + L_2 + \cdots + L_p$.
$\mathbf{Z}_j$	is an $m \times n$ matrix associated with the jth categorical variable. The n columns of $\mathbf{Z}$ refer to the L_j category level points (CLPs), each repeated wherever it appears in the ith sample. Thus there are only L_j distinct columns of $\mathbf{Z}_j$. The m rows of $\mathbf{Z}$ refer to the dimension of the space containing the CLPs. In what follows m takes on the value $n-1$, $K-1$ or $r < m$ depending on the context.
$\mathbf{1}$	is a column-vector of units, whose length may be indicated by a suffix.
$d^2_{ii'}$	is the squared distance between cases i and i'. It is assumed that the distance is additive satisfying $d^2_{ii'} = \sum_{j=1}^{p} f\left(x_{ij}, x_{i'j}\right)$.
$-\frac{1}{2}d^2_{ii'}$	is termed *ddistance* as an abbreviation for $-1/2$ times the squared distance.
$\mathbf{D} \equiv \{-\frac{1}{2}d^2_{ii'}\}$	is an $n \times n$ *ddistance* matrix (not necessarily Pythagorean) generated between the rows of $\mathbf{X}$.
$\mathbf{D}_j \equiv \{-\frac{1}{2}f(x_{ij}, x_{i'j})\}$	is an $n \times n$ *ddistance* matrix (not necessarily Pythagorean) generated by the n rows and jth variable (column) of $\mathbf{X}$. Because distances are additive it follows that $\mathbf{D} = \sum_{j=1}^{p} \mathbf{D}_j$.
$\mathbf{e}_i$	denotes an n-vector with its ith element equal to unity, else zero.

2. Review of biplots for categorical variables in the ungrouped case

In this section we summarise well-known results available e.g. in [9]. These are needed as the starting point for the $K \geq 2$ group extensions discussed in Section 3 with which the results of Section 2 may be compared.