Quality & Quantity (2005) 39:359–372
DOI 10.1007/s11135-005-3005-1

Multivariate Prediction with Nonlinear Principal Components Analysis: Theory[†]

JOHN C. GOWER[1] and JÖRG BLASIUS[2,*]
[1]*Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, U.K.;*
[2]*Seminar for Sociology, University of Bonn, Lennéstr. 27, 53113 Bonn, Germany*

Abstract. We propose the notion of multivariate predictability as a measure of goodness-of-fit in data reduction techniques which are useful for visualizing and screening data. For quantitative variables this leads to the usual sums-of-squares and variance accounted for criteria. For categorical variables we show how to predict the category-levels of all variables associated with every point (case). The proportion of predictions which agree with the true categories gives the measure of fit. The ideas are very general; as an illustration we use nonlinear principal components analysis (NLPCA) in association with ordered categorical variables. A detailed example using data from the International Social Survey Program (ISSP) will be given in Blasius and Gower (quality and quantity, 39, to appear). It will be shown that the predictability criterion suggests that the fits are rather better than is indicated by "percentage of variance accounted for".

Key words: biplot, large scale data analysis, nonlinear principal components analysis, prediction

1. Introduction

Principal components analysis (PCA) is a popular way of displaying relationships between cases associated with several variables. With categorical variables, PCA is not immediately available, although it is not unknown for numerically coded categories to be treated as if they were numerical ratio-scales. The more appropriate analysis is to use nonlinear principal components analysis (NLPCA) in which the categories are replaced by optimal scores (see e.g., Gifi, 1990; Heiser and Meulman, 1994). The optimal scoring

[†] This article was written while John Gower was a visiting professor at the ZA-Eurolab, at the Zentralarchiv für Empirische Sozialforschung, University of Cologne, Germany. The ZA is a Large Scale Facility funded by the Training and Mobility of Researchers program of the European Union.

[*] Author for correspondence: Jörg Blasius, Seminar for Sociology, University of Bonn, Lennéstr. 27, 53113 Bonn, Germany. E-mail: jblasius@uni-bonn.de

process allows order-constraints to be imposed so that ordered categorical variables get increasing, or at least nondecreasing, scores as the category-levels become increasingly severe. When the responses are not consistent with the implied ordering, this manifests itself by giving tied optimal scores for two or more categories. Unlike classical PCA, the number r of dimensions required in the fit must be specified in advance and the solutions for r and $r+1$ dimensions are not nested. Once the optimal scores have been found, they may replace the category codes and the remainder of the analysis may be regarded as a classical PCA. This means that the degree of fit is judged as the ratio of sums-of-squares in the fitted dimensions to the total, usually, the number of categorical variables. Indeed, the NLPCA criterion for estimating the optimal scores is to maximize this ratio for the given value of r.

All the above is well-known. Here we suggest an alternative criterion for judging the fit of any multidimensional scaling (MDS) analysis of categorical variables and exemplify its use in conjunction with NLPCA. As is common, but not universal practice, we subsume methods such as PCA, NLPCA and multiple correspondence analysis among MDS techniques, noting that PCA may be viewed as a special case of classical scaling (e.g. Gower, 1966).

We begin with a short recapitulation of those properties of PCA that are important for our development. In the PCA of a numerical data-matrix $\mathbf{X}$ one obtains an r-dimensional approximation $\hat{\mathbf{X}}$ that minimizes the sum-of-squares $\|\mathbf{X}-\hat{\mathbf{X}}\|$. The solution is given by Eckart and Young (1936) based on the singular value decomposition of $\mathbf{X}$ but is usually presented, equivalently, in terms of the eigenvectors of $\mathbf{X}'\mathbf{X}$, often normalized to be a correlation matrix. The essential thing is that $\hat{\mathbf{X}}$ approximates $\mathbf{X}$ and may be regarded as a surrogate for $\mathbf{X}$ itself. This surrogate is useful when r is small, especially when $r=2$, and the approximation is a good one. Then, $\hat{\mathbf{X}}$ presents a mass of multivariate data in a manageable form that allows ready visual inspection of aspects of the data. We say that $\hat{\mathbf{X}}$ *predicts* $\mathbf{X}$. Note that we are not trying to predict an unknown value that might be observed in the current or some future data-set but to predict known data-values from an r-dimensional approximation. This use of the word *prediction* may be perceived as conflicting with its use in other statistical contexts and to be superfluous to *approximate*. Indeed, *approximate* is acceptable for numerical variables but, in the following, where we are concerned with categorical variables, *predict* is the more appropriate, e.g. it would be inappropriate to say that the color "red" approximates "blue". Perhaps it would be clearer to distinguish the two uses of *prediction* by referring to *internal prediction* (for our usage) and *external prediction* (classical usage) but we think this terminology to be too cumbersome and pedantic. In multivariate analysis, *internal prediction* is nearly always used, either overtly or implicitly, as a preliminary step in data reduction; without it, it would be difficult to make further progress.

The idea developed below is to associate a set of *predicted* category values $\hat{\mathbf{X}}$ with every point in any MDS. The hope is that these predictions agree well with actual category-levels $\mathbf{X}$ in the data, thus giving a surrogate set of categorical values that has lost little information. For the present we assume that $\hat{\mathbf{X}}$ can be computed and discuss how this is done in Section 2. The sum-of-squares criterion $\|\mathbf{X}-\hat{\mathbf{X}}\|$ is meaningless for categorical variables, so it must be replaced. An obvious measure to use is the proportion, or percentage, of correct predictions which might be written $>\mathbf{X}\circ\hat{\mathbf{X}}<$ to emphasize the link with $\|\mathbf{X}-\hat{\mathbf{X}}\|$. In $>\mathbf{X}\circ\hat{\mathbf{X}}<$, $\circ$ represents assessment of agreement between corresponding elements of $\mathbf{X}$ and $\hat{\mathbf{X}}$ while $><$ denotes that the number of agreements is summed.

The difference between evaluating sums-of-squares and counting agreements, is less than it may seem. The key to both approaches, is the measure of the distance of a point O, say, from a set of points. When the set of points form a continuum, such as a line, plane or r-dimensional subspace, whose points are labeled by some numerical co-ordinate system, then the shortest distance is given by the orthogonal projection of O onto the set, together with its numerical label; this is the situation for quantitative variables. With a finite set of n nominally labeled points, the shortest distance is merely the shortest of the n distances from O together with its associated nominal label. As is explained below, the latter, less stringent criterion, is especially appropriate for use with the discrete nature of categorical variables. In both cases, one is finding the nearest acceptable value, numerical or categorical, as the case may be.

In NLPCA and similar methods, the optimal scoring process replaces the categories by a *discrete* set of numerical values, thus transforming the problem into a conventional PCA. The distinction between a discrete set of quantified nominal values and a continuum of values on a scale is a potential source of misconceptions. In the following development, we attempt to respect the discrete nature of the data as much as possible. The prediction criterion gives a more natural measure of the degree of approximation to categorical information than does variance accounted for. Ideally, we would like an MDS method that *optimizes* the criterion $>\mathbf{X}\circ\hat{\mathbf{X}}<$ in r dimensions, but this seems to be a hard problem. Failing that we can see how well a method such as NLPCA performs as judged against the new criterion.

In the second part of this paper (Blasius and Gower, to appear) we will illustrate our ideas using data drawn from the 1995 International Social Survey Program (ISSP) and compare the traditional fit criterion *variance accounted for* with the new predictability criterion.

2. Methodology

2.1. PREDICTING CATEGORICAL VARIABLE LEVELS

We first outline the basic ideas of prediction in general terms and then show how they relate to NLPCA. Gower and Hand (1996) as well as Gower and Harding (1998) introduced the idea of representing the levels of categorical variables by *neighbor regions* approximated in MDS representations by *prediction regions*. We illustrate their ideas with artificial data (see Figure 1).

Figure 1 shows prediction regions for a categorical variable *Color*, with four levels (or categories). The numbers refer to 25 cases of which 1–6 are *red*, 7–11 *green*, 12–16 *yellow* and 17–25 *blue*. The 25 cases are plotted, say by some form of MDS, together with the four prediction regions. Each case is labeled by the name of its true color.[1] The cases that fall within the region labeled *red* are predicted to be that color and so on for the other colors. Numbers 1–6 are all *red* and 1–4 are plotted correctly in the "red region", but number five is plotted in the "yellow region" and number six in the "blue region", these last two are incorrectly predicted. Table I lists the predicted against the true values for all the colors.

From Table I we see that the correct predictions occur on the diagonal and sum to $14=4+3+3+4$. The remaining 11 cases occur off the diagonal and give incorrect predictions. Thus, Figure 1 represents 56% correct and 44% incorrect predictions.

Diagrams such as Figure 1 occur naturally in many representations of cases described by categorical variables. They are associated with the existence of category level points (CLPs). A categorical variable with L_k levels has L_k CLPs, each on a different axis, which for the present we shall assume to be mutually orthogonal. In our example where the kth variable is *Color*, $L_k=4$ and the four CLPs refer to the four levels. The most simple

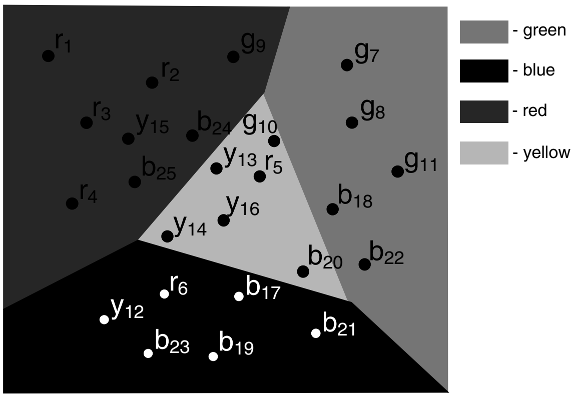

Figure 1. Prediction regions for a variable *Color* with four levels.

Table I. The numbers of correct and incorrect predictions associated with Figure 1

	Predictions				
True	Red	Green	Yellow	Blue	Sum
Red	4	0	1	1	6
Green	1	3	1	0	5
Yellow	1	0	3	1	5
Blue	2	2	1	4	9
Sum	8	5	6	6	25

choice of co-ordinates for the CLPs is (1,0,0,0) for *red*, (0,1,0,0) for *green*, (0,0,1,0) for *yellow*, and (0,0,0,1) for *blue*; we shall meet other settings below. Whatever the actual co-ordinate settings for the CLPs, each has an associated *neighbor-region*. The one for *red* consists of all points that are nearer the CLP for *red* than they are to any other CLP and normally will contain all cases that are *red*; similarly for the neighbor-regions associated with the other colors. These neighbor-regions are convex. Similar convex neighbor-regions exist for other categorical variables but with their CLPs in orthogonal higher dimensional spaces. With p categorical variables and a total of $L=\sum_{k=1}^{p} L_k$ category levels, this requires an L dimensional space and, to continue with the simple choice of CLPs used above for *Color*, the whole set of CLPs is given by the rows of the $L \times L$ unit matrix $\boldsymbol{I}$. More generally, the CLPs have co-ordinates given in a diagonal matrix $\boldsymbol{D}$. For example, in multiple correspondence analysis $\boldsymbol{D}=\boldsymbol{L}^{-1/2}$ where $\boldsymbol{L}=\mathrm{diag}(\boldsymbol{L}_1,\boldsymbol{L}_2,\ldots,\boldsymbol{L}_p)$ and $\boldsymbol{L}_k$ is an $\boldsymbol{L}_k \times \boldsymbol{L}_k$ diagonal matrix giving the frequencies of occurrence for the levels of the kth variable; for our variable *Color* these frequencies take the values 6, 5, 5, and 9, the row sums of Table I.

When $\boldsymbol{G}$ is an $n \times L$ indicator matrix, the rows of $\boldsymbol{GD}$ give the co-ordinates of all the n cases; every case falls into its correct neighbor-region. Thus, all cases that are *red* and no cases that are not *red* fall into the neighbor-region for *red*. Values of quantitative variables to be associated with a point (usually representing a case) are found by projecting onto scaled co-ordinate axes and reading off the correct values. More fundamental than projection is that one is reading off, on each axis, the *nearest* scale marker to the point. Correspondingly, for categorical variables one reads off the label associated with the nearest CLP. The simplex of CLPs associated with a categorical variable may be viewed as a generalization of the concept of a co-ordinate axis for a quantitative variable, an idea developed further by Gower (2002).

The set-up just described uses L dimensions. The process of predicting values of variables associated with points in an r-dimensional approximation (usually, $r = 2$), using only information in that space, is the concern of biplots. For quantitative variables and in linear cases biplots are expressed as vectors and the approximation is often interpreted through singular value decompositions (Gabriel, 1971, 1981; Greenacre, 1993). Although mathematically valid, this approach can be simplified; Gower and Hand (1996) stress that biplot axes may be interpreted just like other co-ordinate axes (i.e. by projection and reading a value off a scale). For categorical variables the process is in some ways more simple and in others more complicated. It is simple because the prediction of a category level to be associated with any point is merely a matter of deciding in which neighbor-region the point lies, just as in Figure 1. This is easily calculated but one also needs a nice visualization. In principle, this is merely a matter of showing how the neighbor-regions intersect the r-dimensional space to give a tessellation of convex *prediction-regions*, again as in Figure 1. Computational procedures for doing this are described in Gower (1993) and in Gower and Hand (1996).[2]

With linear biplots, each quantitative variable has one axis, so all p axes are conveniently shown on a single diagram. However, each categorical variable generates one set of prediction regions, as in Figure 1. Usually it would be too confusing to superimpose the regions for two or more categorical variables on the same diagram (see Gower and Hand, 1996, for an example). The prediction-regions for each variable are best shown separately; how they overlap, which relates to correlation, may be perceived through mental visualization or perhaps by stacking transparencies. In the next section we show how for the case of *ordered* categorical variables, all variables may be shown simultaneously, as in the linear case.

2.2. PREDICTION REGIONS FOR NONLINEAR PRINCIPAL COMPONENTS ANALYSIS

It is convenient to develop our discussion of the form taken by CLPs and prediction regions for ordered categorical variables in the context of NLPCA. Initially, we assume that levels are unordered. So far, we have assumed that the CLPs have orthogonal co-ordinate representations on L axes and that $\boldsymbol{D}$ is diagonal. This restriction is not necessary. For example, in NLPCA $\boldsymbol{D} = \text{diag}(z_1, z_2, \ldots, z_p)$ where z_k is a column-vector of length L_k, giving the optimal quantifications for the kth variable, so that in this case $\boldsymbol{D}$ has dimensions $L \times p$. With (ordered or unordered) categorical data coded as a binary indicator matrix $\boldsymbol{G}$, $\boldsymbol{GD}$ converts the data to numerical scores, giving an $n \times p$ matrix which may be analyzed by conventional PCA, though most NLPCA computer output provides this automatically (see SPSS, 1999, CATPCA, formerly PRINCALS).

In line with conventional PCA, the quantified variables are scaled to have zero mean and normalized to have unit sum-of-squares, thus $\sum_{i=1}^{L_k} l_k i z_{ki} = 0$ and $\sum_{i=1}^{L_k} l_{ki} z_{ki}^2 = 1$, or in matrix notation $\mathbf{1}'\boldsymbol{L_k Z_k} = 0$ and $z_k'\boldsymbol{L}_k z_k = 1$, for $k = 1, 2, \ldots, p$. With this scaling of the scores, the principal components are the eigenvectors of a correlation matrix $\mathbf{D}'\mathbf{LD}$ with trace p. In NLPCA the scores are chosen to give a best PCA in some *specified* number, r, of dimensions (in our case $r = 2$) by maximizing the sum of the r principal eigenvalues relative to the total trace, p. Because the CLPs for a single variable are given in a vector z_k (the quantifications), it follows that they have a linear representation even though the category-levels are unordered. Therefore the CLPs for the kth variable are L_k points lying on a line ξ_k, say, through the origin. The CLPs are placed distances $z_{k1}, z_{k2}, \ldots$ from the origin. This gives what looks very much like a conventional co-ordinate axis. However, the similarity is deceptive because intermediate values are undefined (e.g. what meaning is there to saying that a person is between *single* and *divorced?*). It would be misleading to project onto such a *pseudo-linear* axis and read off scale values.

In the example to be discussed in Blasius and Gower (to appear), where the category-levels are *ordered*, the CLPs for the kth variable are also ordered, to give something looking even more like a conventional co-ordinate axis. Nevertheless, values intermediate to the CLPs remain undefined. The geometrical setup is shown in Figure 2.

Assume a four-point variable with ordered categories: N (none), L (little), S (some) and M (much). In Figure 2, the ordered category quantifications are shown on an axis ξ_k. If we label the plane holding the r-dimensional display L_r, the normal planes at points half way between the category quantifications defining the CLPs are the boundaries of the neighbor-regions. These intersect with L_r to give the shaded prediction-regions bounded by a series of parallel lines, as shown in Figure 2. This would complete the calculation for the more general convex prediction-regions,

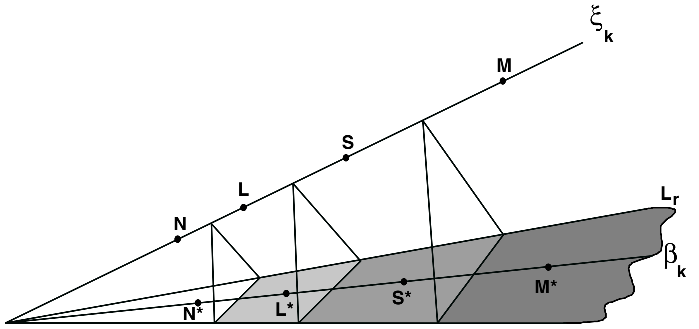

Figure 2. Category level points for an ordered categorical variable.

as shown in Figure 1, that arise from nonlinear CLPs. In the important case of pseudo-linearity induced by ordinal variables a great simplification is possible. Then, we may replace the prediction regions for a variable by a single line β_k orthogonal to the intersections, marking the positions of new CLP*s (N*, L*, S* and M* in Figure 2) in the approximation space L_r. It is then only necessary to decide which CLP* is nearest a point representing a case, to decide in which prediction region it lies. The great advantage of this representation is that it allows the variables to be shown as a set of intersecting lines $\beta_1, \beta_2, \ldots, \beta_p$ just as for linear biplots, thus avoiding the difficulties of superimposing sets of convex prediction regions that occur with the full generality of CLPs permitted for categorical variables. When an order constraint has been applied for the calculation of the CLPs, they will not only be linear but also properly ordered as is shown in Figure 2. In the case of ties, say when L and S have the same quantifications, one may make an infinitesimal perturbation between both, thus keeping the original order of the levels and predicting L for points to the left of L*/S* and S for points to the right. It might be argued that this is not appropriate and that in reality one cannot distinguish between predicting either L or S. Then, if the original data gave either an L or an S, one would have a correct prediction and, if not, prediction would be incorrect. Of course, when there are three or more ties, the central categories are never predicted. When there is preponderance of ties, the indication is that the data are not consistent with ordinality and then one should not be imposing ordinal constraints.

Note that the SVD plays no part in the interpretation and that the plotted points CLP* are not projections of the true CLPs onto the PCA space, indeed they are *back projections* (see Gower and Hand, 1996). The ability of providing CLP*s within L_r itself is a consequence of the linearity of the CLPs and is not a property that extends to nonlinearly arranged CLPs. The use of biplot axes for NLPCP is illustrated in Figure 3 which shows two variables "1" and "2" each described by four ordered categories (N = *none*, L = *little*, S = *some*, M = *much*).

The prediction for a case R by nearness to the CLPs is shown in Figure 3 and contrasted with projection. Things to notice are:

(i) each axis stops at the extremes of each scale, thus emphasizing the meaninglessness of points beyond,

(ii) there are no horizontal and vertical axes because these would add nothing to the information given by the biplot axes,[3]

(iii) the predictions (S_1, L_2) for R by nearness to the CLP*s can also be achieved by projecting onto the axes and finding the CLP*s nearest the projections, and

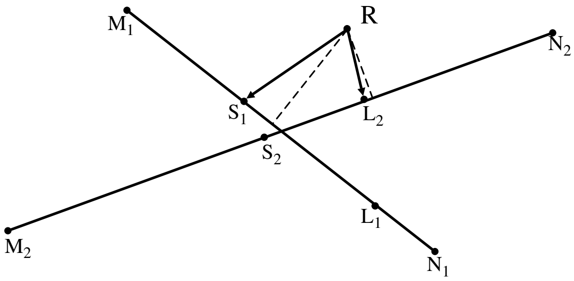

Figure 3. Prediction with two ordered categorical variables "1" and "2". Variable i has CLP*s N_i, L_i, S_i and M_i. The arrowed lines give the nearest CLP*s predicted for the case labeled R.

(iv) that the projections given by the dashed lines differ from the nearest CLP*s given by arrowed lines.

Regarding (i) we might add that the lines between the extremes should also be suppressed. Not only would this discourage interpretation of projected points, but it would do so without impairing nearness interpretations. On balance, we feel that the lines are justified as a visual aid in linking successive ordinal levels.

As discussed above, prediction regions are not shown because with ordered categorical variables they are bounded by lines as shown in Figure 2. The representation of ordered categorical variables has characteristics both of quantitative and of unordered categorical variables; the linearity of axes (here pseudo-linearity) is retained but scale markers are replaced by category names.

To sum up: with quantitative variables we predict the responses of an individual by projecting onto the different axes and reading off the scale values of the variables. This procedure is valid for both the original co-ordinate axes and the biplot axes. We have seen that with unordered categorical variables things are quite different but with ordered categorical variables we come close to the classical usage. The category levels appear on linear axes but the intervening points do not have the associated scale values required to justify projection; projection may be replaced by nearness to CLP*s.

In practice, to calculate the proportion of correct and incorrect predictions as in Table I we need to compute how close is each case in the approximation space L_r to the true CLPs. The details of how to do this are given in Appendix A.

2.3. SOME GENERAL REMARKS

To close this section, we add a few remarks about the more general uses of CLPs. One dimensional sets of CLPs, as for NLPCA, and sets occupying

L_k dimensions whose co-ordinates are represented in a diagonal matrix, are two extreme types of representation; many intermediates are possible. For example, we could allow $\boldsymbol{L}_k$ to be a general $L_k \times L_k$ matrix, in which case the L_k CLPs may be regarded as occurring on oblique axes, i.e. they form a simplex as in Figure 4b rather than the simplex shown in Figure 4a.

Also, in NLPCA every variable can be replaced by two or more dimensions, so that the vector z_k becomes a matrix $\boldsymbol{Z}_k$ with two or more columns, the so-called *multiple quantifications* of CATPCA (see Gifi, 1990, SPSS, 1999). Indeed, it would be sensible to use multiple quantifications with unordered categorical variables, since otherwise a spurious ordering may be implied by the pseudo-linearity of optimal scores set along a single axis. In these representations, the axes for all pairs of variables are assumed to lie in mutually orthogonal subspaces, but even this restriction could be dispensed with by allowing the CLPs of two variables to share part of one another's space. However, this degree of generality destroys the *nearness* properties that are important for interpretation.

Finally, we draw attention to Guttman's ideas on facet theory which bear some relation to our proposals (see, for example, Guttman, 1965; Borg and Shye, 1995; Borg and Groenen, 1997). The regional interpretations in facet theory are associated with an MDS diagram in any convenient way that has a substantive interpretation in terms of the original variables. Boundaries are allowed to have any shape determined in the light of the user's judgment, rather than by any mathematical methodology. The predictions we make above are associated with prediction-regions, which may be regarded as a species of facet for ordered categorical variables. Our prediction-regions are mathematically derived and are bounded by parallel lines. With unordered categorical variables, the prediction regions become convex

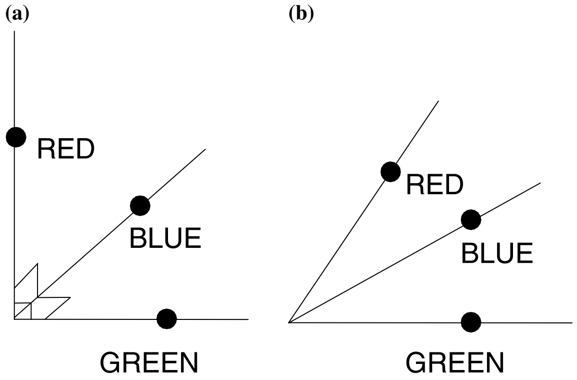

Figure 4. CLPs for a categorical variable with three levels, (a) relative to orthogonal axes giving a diagonal matrix $\boldsymbol{L}$ and (b) relative to oblique axes giving a general 3×3 matrix $\boldsymbol{L}$.

figures with linear boundaries (see Gower and Hand, 1996; Gower and Harding, 1998). Thus, our regions are more restrictive than the generality allowed by facet theory but they are more objective and are associated with an intuitively appealing criterion.

In the second part of the paper (Blasius and Gower, to appear) we will give an example using data from the ISSP to demonstrate how the proposed procedure works and to compare our results with those retained from NLPCA and classical PCA. Comparing the solutions will also allow for some comments on the quality of data from the participating countries.

3. Discussion

We have proposed a measure of fit based on multivariate predictability. Predictions may be made by finding the nearest one of a set of CLPs to each case represented in a low-dimensional representation. The set of predictions may then be compared with the values given in the raw data, giving a measure of the number of correct predictions to be associated with the low-dimensional representation. The ideas have been set in the context of ordered categorical variables in NLPCA, but can be readily extended to most other forms of MDS and to other kinds of variable. Numerical and graphical illustrations using two sets of questions with ordinal response scales, drawn from 23 countries participating in the 1995 ISSP will be given in Blasius and Gower (to appear).

In this paper we will show that when the quality of the data is good, *prediction* performs well. When they are bad, as judged in NLPCA by recourse to many ties, *prediction* can be poor. In contrast, *variance accounted* for will give some good fits to bad data and some bad fits to good data. Further, the notion of (internal) predictability has intuitive appeal and is more directly related to what is required of a surrogate approximation to the data than are criteria based on *variance accounted for*. The predictability criterion itself may be used as a basis for defining new forms of analysis (Gower, 2002).

Appendix A: The Computation of Predictions with Special Reference to CATPCA in SPSS

In the CATEGORIES software of SPSS (1999) one can select the option *optimal scaling* and using the option *some variables are not multiple nominal* select CATPCA (categorical principal components analysis; in earlier versions PRINCALS). In our case all variables are treated as *ordinal with single quantifications*. For each variable, CATPCA gives the quantifications of all levels z_{ki} where k refers to the variable and i to the level of the variable.

We have chosen a two-dimensional space to represent the cases (e.g. the respondents).

The output is not immediately in the form we require, it needs some additional calculations as described in the following. We shall be concerned with the normalizations of the quantified scores and with the normalizations of the eigenvalues and eigenvectors.

The scores are already normalized so that $\mathbf{1}'\boldsymbol{L}_k z_k = 0$ and $z_k'\boldsymbol{L}_k z_k = 1$, $k = 1, \ldots, p$. This means that the columns of the matrix of quantifications $\mathbf{Z}$ have zero means and unit variances, so that $\mathbf{Z}'\mathbf{Z}$ is a correlation matrix. The fitted two-dimensional co-ordinates $\mathbf{X}(n \times 2)$ of the respondents are also immediately available.

For a conventional PCA we need the eigenvalues and eigenvectors satisfying $(\mathbf{Z}'\mathbf{Z})\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$ to be normalized so that $\mathbf{V}'\mathbf{V} = I$ and $\sum_{k=1}^{p} \lambda_k = p$. CATPCA currently offers several different normalizations of eigenvectors $\mathbf{V}_P$ (referred to as *component loadings*), and some small adjustments may be needed to ensure the above normalizations, before proceeding with the calculations described below.

We have to compute the distances of every case, as represented in two dimensions, from each of the CLPs. Each variable has four collinear CLPs. The occurrence of ties needs special attention when calculating predictions. When two levels of a variable have tied quantifications, their CLPs coincide, so there is ambiguity over which level to predict. We have resolved this by making a small perturbation to the adjacent CLPs. The consequence is that all points to the "right" will be predicted with the higher level and all points to the "left" will be predicted with the lower level. With three or more tied levels, we do a similar perturbation and then the central level(s) will hardly ever be predicted. In this way the separation of the CLPs is ensured. The geometry is shown in Figure 5.

In Figure 5, C is the CLP for a quantification ζ on the kth variable. C has co-ordinate values $\zeta\mathbf{e}_k$ (where $\mathbf{e}_k$ is a unit row-vector with one in its kth position and zero elsewhere). X represents one case with co-ordinates (x_1, x_2) in its two-dimensional representation; the values (x_1, x_2) are supplied by the program and can be stored along with the original data. Y is the projection of C onto L_r. Let us denote the loadings of the variables on the two axes by v_{kr} (k is the variable and $r = 1, 2$, the dimension) and by $\mathbf{V}_2$ the first two columns of $\mathbf{V}$. We require the distance d_{CX}, which from Figure 5 we see is given by:

$$d_{\mathrm{CX}}^2 = d_{\mathrm{CY}}^2 + d_{\mathrm{XY}}^2.$$

Thus, to calculate the distance d_{CX} requires the following steps:

(i) Calculate the projection Y of C onto the plane of approximation. This is given by $\mathbf{y} = \zeta\mathbf{e}_k\mathbf{V}_2$. This operation picks out the kth row of

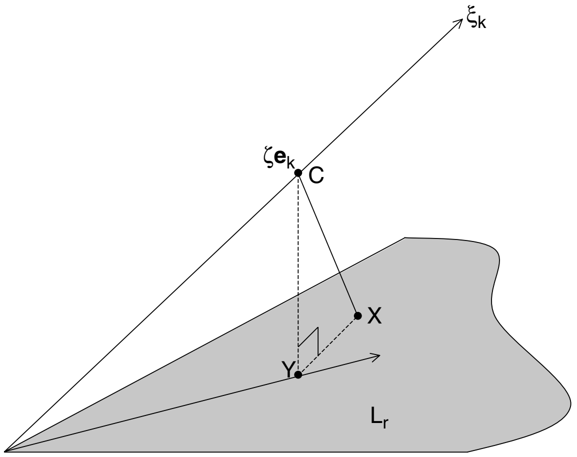

Figure 5. The geometry of prediction

$\mathbf{V}_2$ and multiplies it by ζ. Thus, $\mathbf{y} = (y_1, y_2)$ are the coordinates of Y.

(ii) Next $d^2_{\mathrm{CY}} = \zeta^2 - d^2_{\mathrm{OY}} = \zeta^2 - (y_1^2 + y_2^2)$.

(iii) $d^2_{\mathrm{XY}} = (x_1 - y_1)^2 - (x_2 - y_2)^2$ so that the required squared distance is given by $d^2_{CX} = (x_1 - y_1)^2 - (x_2 - y_2)^2 + \zeta^2 - (y_1^2 + y_2^2)$ which simplifies to $\left(x_1^2 + x_2^2\right) - 2\,(x_1 y_1 + x_2 y_2) + \zeta^2$.

(iv) The quantity obtained in step (ii) should be computed for each value (z_{k1}, z_{k2}, z_{k3} and z_{k4}) of ζ corresponding to the four quantifications of the kth variable. The term $\left(x_1^2 + x_2^2\right)$ in step (iii) remains constant so may be ignored. The value of z_{ki} giving the smallest distance, predicts the level i for the kth variable, for the case denoted by X.

Steps (i) to (iv) have to be done for all p variables ($k = 1, \ldots, p$) and for all n cases X, finally giving an ($n \times p$) matrix of predictions. Simple cross-tabulation of the prediction variables with the original variables provides the number of correct and incorrect predictions, as in Table I.

The biplot axis for the kth variable is essentially computed at step (i), above. As ζ varies (y_1, y_2) trace out the biplot axis and markers may be shown at each of the four quantification values (z_{k1}, z_{k2}, z_{k3} and z_{k4}). Remember, that these calculations must be done using the properly scaled values of $\mathbf{V}$, $\boldsymbol{\Lambda}$; this means that special attention has to be paid to ensure that the output of CATPCA is in the required form. Any rescaling needed may be done either at the outset or incorporated into the formulae given in steps (i)–(iv).

Notes

[1] An improved display would be to color the cases and the prediction regions. Then, cases which do not match their background would show-up better and cases which do match their background would show as small open circles (see e.g. Gower and Harding, 1998).

2. A program in Genstat (see Genstat 5 Committee, 1993; Payne et al., 1998) has been written by Simon Harding, another one in S-plus (see Chambers, 1998) by Roger Ngouenet.
3. This point applies to any factorial type representation relative to principal axes.

References

Blasius, J. & Gower, J. C. (to appear). Multivariate prediction with nonlinear principal components analysis: Application. *Quality and Quantity* 39.

Borg, I. & Groenen, P. (1997). *Modern Multidimensional Scaling*. New York: Springer.

Borg, I. & Shye, S. (1995). *Facet Theory. Form and Content*. Newbury Park, CA: Sage.

Chambers, J. M. (1998). *Programming with Data: A Guide to the S Language.* New York: Springer.

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* 1: 211–218.

Gabriel, K. R. (1971). The biplot-graphic display of matrices with applications to principal components analysis. *Biometrika* 58: 453–467.

Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspecting of data and diagnosis. In: V. Barnett (eds.), *Interpreting Multivariate Data*, Chichester: Wiley, pp. 147–174.

Genstat 5 Committee. (1993). *Genstat 5 Release 3 Reference Manual*. Oxford: Numerical Algorithms Group.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester: Wiley.

Gower, J. C. (1966): Some distance properties of latent-root and vector methods used in multivariate analysis. *Biometrika* 53: 325–338.

Gower, J. C. (1993). The construction of neighbour-regions in two dimensions for prediction with multi-level categorical variables. In: O. Opitz, B. Lausen & R. Klar (eds.), *Information and Classification: Concepts–Methods–Applications Proceedings 16th Annual Conference of the Gesellschaft für Klassifikation*, *Dortmund, April 1992*, Berlin: Springer, pp. 174–189.

Gower, J. C. (2002). Categories and quantities. In: S. Nishisato, Y. Baba, H. Bozdogan & K. Kamefuji (eds.), *Measurement and Multivariate Analysis*, Tokyo: Springer, pp. 1–12.

Gower, J. C. & Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.

Gower, J. C. & Harding, S. (1998). Prediction regions for categorical variables. In: J. Blasius & M. Greenacre (eds.), *Visualization of Categorical Data*. San Diego: Academic Press, pp. 405–423.

Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics* 20: 251–269.

Guttman, L. (1965). A faceted definition of intelligence. *Scripta Hierosolymitana* 14: 166–181.

Heiser, W. J. & Meulman, J. J. (1994). Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In: M. J. Greenacre & J. Blasius (eds.), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*, London: Academic Press, pp. 179–209.

Meulman, J. J. & Heiser, W. J. (1999). *SPSS Categories 10.0*. Chicago: SPSS Inc.

Payne, R. W., Lane, P. W., Baird, D. B., Gilmour, A. R., Harding, S. A., Morgan, G. W. Murray, D. A., Thompson, R., Todd, A. D., Tunicliffe-Wilson, G., Webster, R. & Welham, S. J. (1998). *Genstat 5 Release 4.1 Reference Manual Supplement*. Oxford: Numerical Algorithms Group.

SPSS. (1999). See Meulman and Heiser (1999).