COMPARISON OF DIMENSIONALITY REDUCTION METHODS APPLIED TO ORDINAL DATA

Martin Prokop – Hana Řezanková

Abstract

The paper deals with the comparison of data dimensionality reduction methods with emphasis on ordinal data. Categorical and especially ordinal data we frequently obtain from questionnaire surveys. A questionnaire usually includes a big amount of questions (variables). For applications of multivariate statistical methods, it is useful to reduce the number of these questions and create new latent variables, which represent groups of original questions. Some dimensionality reduction methods are applicable to ordinal data (latent class models), some methods must be improved (categorical principal component analysis). Other methods are based on a distance matrix, so it is possible to use an appropriate distance measure for ordinal data (multidimensional scaling). In this paper, dimensionality reduction methods are applied to real datasets including ordinal data in the form of Likert scales. Various techniques for the comparison of these methods are used. They are aimed to investigate goodness of the data structure in original and reduced space. In this paper the goodness is evaluated by Spearman rank correlation coefficient.

Key words: dimension reduction, principal component analysis, multidimensional scaling, latent class models

JEL Code: C3, C6, C8

Introduction

The aim of this paper is the comparison of dimensionality reduction methods for ordinal variables. Reduction methods described in the following chapter were applied to ordinal datasets including values in the form of Likert scales. Inter-object distances in original and reduced space were evaluated. With respect to the ordinal character of the data, Kendall correlation coefficient was used as a similarity measure. Further we measured how well the structure and structural relationship of the data were preserved by dimension reduction. For the purpose of this paper Spearman rank correlation coefficient between inter-object distances in original and reduced space was used. In the current research a similar problem but for the

different kind of the data (continuous) or for different kinds of methods (nonlinear) was solved in (Li, 1995). For the comparison various procedures were used, e.g. scatterplots or Spearman rank correlation coefficient of inter-object distances in original and reduced space, Procrustes analysis or measuring the generalization error of *k*-nearest neighbor classifiers trained on the resulting data representations (Maaten, 2008). Similar reduction methods applied to ordinal data evaluated by fuzzy cluster analysis were discussed in (Sobíšek, 2011).

1 Dimensionality reduction methods

Basic methods of the data dimensionality reduction are principal component analysis (PCA), factor analysis (FA) and multidimensional scaling (MDS). Classical FA methods assume linear relations among original variables, new latent variables are continuous and normally distributed. Conventional factor analysis is usually based on correlation matrix analysis, for more details see (Hebák et al., 2007).

Common methods of latent variables identification are latent class models. There exist various methods, which are available in statistical software packages, e.g. basic LCA models, latent class cluster models LCC, discrete factor analysis models DFactor, latent trait analysis LTA, latent profile analysis LPA, latent class regression models LCR etc.

1.1 Categorical principal component analysis

Some methods are based on multidimensional space projection into the space with lower dimension. A basic method is principal component analysis. The aim is to find a real dimension of the data. To find a real dimensionality, original dataset **X** is transformed to the new coordinate system by an orthogonal linear transformation. Let F_s (resp. G_s) be the vector of the rows coordinates (resp. columns) on the axis on the *s*-th rank. These two vectors are related by the transition formula, e. g. in the case of PCA there are

$$F_S(i) = \frac{1}{\sqrt{\lambda_S}} \sum_k x_{ik} m_k G_S(k), \tag{1}$$

$$G_S(k) = \frac{1}{\sqrt{\lambda_S}} \sum_k x_{ik} p_i G_S(i), \tag{2}$$

where F_s denotes the coordinate of the *i*-th object on the *s*-th axis, G_s denotes the coordinate of the *k*-th variable k on the *s*-th axis, λ_s the eigenvalue associated with the *s*-th axis, m_k is the weight associated to the *k*-th variable , p_i is the weight associated to the *i*-th object.

Instead of conventional principal component analysis for quantitative variables it is possible to use categorical principal component analysis CATPCA, which transforms categorical variables into quantitative variables and does not assume linear relations among variables. For more datails see (Le, 2008).

1.2 Multidimensional scaling

According to (Holland, 2008) this method starts with a matrix of data **X** consisting of *N* rows of objects and *J* columns of variables. From this symmetrical matrix of all pairwise distances among objects is calculated with an appropriate distance measure, such as Euclidean distance, Manhattan distance (city block distance), and Bray distance. The MDS ordination will be performed on this distance matrix. Next, a desired number of *m* dimensions is chosen for the ordination. Distances among objects in the starting configuration are calculated, typically with the Euclidean metric. These distances are regressed against the original distance matrix and the predicted ordination distances for each pair of objects is calculated. A variety of regression methods can be used, including linear, polynomial, and non-parametric approaches. In any case, the regression is fitted by least-squares. The goodness of fit of the regression is measured based on the sum of squared differences between ordination-based distances and the distances predicted by the regression. This goodness of fit is called stress and can be calculated in several ways, e.g. with one of the most common being Kruskal's Stress

$$Stress = \sqrt{\frac{\sum_{h,\,i}\left(d_{hi} - \hat{d}_{hi}\right)^2}{\sum_{h,i} d_{hi}^2}}, \qquad (3)$$

where d_{hi} is the ordinated distance between *h*-th and *i*-th objects, and $\hat{d}$ is the distance predicted from the regression. The basic similarity measure of two quantitative variables is Pearson correlation coefficient. To measure similarity of ordinal variables it is possible to use e.g. Spearman or Kendall rank correlation coeficient or symmetric Sommers coefficient. For details see e.g. (Hendl, 2006).

1.3 Latent class models

The basic latent class model is a finite mixture model, in which the component distributions are assumed to be multi-way cross-classiffcation tables with all variables mutually independent. The latent class regression model further enables us to estimate the effects of covariates on predicting latent class membership. Evaluation algorithm uses expectation-maximization and Newton-Raphson algorithms to find maximum likelihood estimates of the model parameters.

According to (Linzer, 2011) the basic latent class model is a finite mixture model in which the component distributions are assumed to be multi-way cross-classification tables with all variables mutually independent. We observe J polytomous categorical variables (the manifest variables), each of which contains K_j possible outcomes, for objects $i = 1, ..., N$. The manifest variables may have different numbers of outcomes, hence the indexing by j. We denote as Y_{ijk} the observed values of the J manifest variables such that $Y_{ijk} = 1$ if respondent i gives the k-th response to the j-th variable, and $Y_{ijk} = 0$ otherwise, where $j = 1, ..., J$ and $k = 1, ..., K_j$. The latent class model approximates the observed joint distribution of the manifest variables as the weighted sum of a finite number R of constituent cross-classification tables. Let π_{jrk} denote the class-conditional probability, that an object in class $r = 1, ..., R$ produces the k-th outcome on the j-th variable. Within each class, for each manifest variable, therefore $\sum_{k=1}^{K_j} \pi_{jrk} = 1$. Further we denote as p_r the R mixing proportions that provide the weights in the weighted sum of the component tables, with $\sum_r p_r = 1$. The values of p_r are also referred to as the prior probabilities of latent class membership, as they represent the unconditional probability that an object will belong to each class before taking into account the responses Y_{ijk} provided on the manifest variables. The probability that the i-th object in the r-th class produces a particular set of J outcomes on the manifest variables, assuming conditional independence of the outcomes Y given class memberships, is the product

$$f\left(Y_i; \pi_r\right) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left(\pi_{jrk}\right)^{Y_{ijk}} . \qquad (4)$$

The probability density function across all classes is the weighted sum

$$P\left(Y_i \mid \pi, p\right) = \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left(\pi_{jrk}\right)^{Y_{ijk}} . \qquad (5)$$

The parameters estimated by the latent class model are p_r and π_{jrk}. Given estimates $\hat{p}_r$ and $\hat{\pi}_{jrk}$ of p_r and π_{jrk}, respectively, the posterior probability that each object belongs to each class,

conditional on the observed values of the manifest variables, can be calculated using Bayes' formula:

$$\hat{P}(r_i \mid Y_i) = \frac{\hat{p}_r f(Y_i; \hat{\pi}_r)}{\sum_{q=1}^{R} \hat{p}_q f(Y_i; \hat{\pi}_q)}, \tag{6}$$

where $r_i \in \{1, \ldots R\}$. Let us recall that the $\hat{\pi}_{jrk}$ are estimates of outcome probabilities conditional on the *r*-th class. It is important to remain aware that the number of independent parameters estimated by the latent class model increases rapidly with *R, J,* and K_j . Given these values, the number of parameters is $R\sum_j (K_j - 1) + (R - 1)$. If this number exceeds either the total number of objects, or one fewer than the total number of cells in the cross-classification table of the manifest variables, then the latent class model will be unidentified. The poLCA estimates the latent class model by maximizing the log-likelihood function

$$\ln L = \sum_{i=1}^{N} \ln \sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{jrk})^{Y_{ijk}} \tag{7}$$

with respect to p_r and π_{jrk}, using the expectation-maximization (EM) algorithm. This log-likelihood function is identical in form to the standard finite mixture model log-likelihood. As with any finite mixture model, the EM algorithm is applicable because each object's class membership is unknown and may be treated as missing data. The EM algorithm proceeds iteratively. We start with arbitrary initial values of $\hat{p}_r$ and $\hat{\pi}_{jrk}$, and label them $\hat{p}_r^{old}$ and $\hat{\pi}_{jrk}^{old}$. In the expectation step, we calculate the missing class membership probabilities using Equation 6, substituting in $\hat{p}_r^{old}$ and $\hat{\pi}_{jrk}^{old}$. In the maximization step, we update the parameter estimates by maximizing the log-likelihood function given these posterior $\hat{P}(r_i \mid Y_i)$ with

$$\hat{p}_r^{new} = \frac{1}{N} \sum_{i=1}^{N} \hat{P}(r_i \mid Y_i) \tag{8}$$

as the new prior probabilities and

$$\hat{\pi}_{jr}^{new} = \frac{\sum_{i=1}^{N} Y_{ij} \hat{P}(r_i \mid Y_i)}{\sum_{i=1}^{N} \hat{P}(r_i \mid Y_i)} \tag{9}$$

as the new class-conditional outcome probabilities. In Equation 9, $\hat{\pi}_{jr}^{new}$ is the vector of length K_j of class-*r* conditional outcome probabilities for the *j*-th manifest variable; and Y_{ij} is the $N \times K_j$ matrix of observed outcomes Y_{ijk} on that variable. The algorithm repeats these steps,

assigning the new to the old, until the overall log-likelihood reaches a maximum and ceases to increment beyond some arbitrarily small value.

2 Distance and similarity measures for ordinal data

Dimension reduction methods frequently require an appropriate similarity measure. For ordinal variables it is possible to use some dependence intensity measure, e.g. an association measure in contingency tables. According to (Hendl, 2006) generally we measure the dependence intensity of two normally distributed variables by Pearson correlation coefficient. If we do not know distribution of the data instead of Pearson correlation coefficient we can use Spearman rank correlation coefficient (10) which can be used also for ordinal variables. Sometimes we also know only ranks of measured values. If there is a similar rank of two samples *X*, *Y*, it means information about the dependence of these variables. Spearman correlation coefficient is evaluated as a correlation coefficient applied to the rank of the values from ranked samples. Values of the variables *X*, *Y* we rank with respect to the size and we obtain the sequence $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$, $Y_{(1)} \le Y_{(2)} \le ... \le Y_{(n)}$. Let R_i be the rank of the variable X_i and Q_i the rank of the variable Y_i in the ranked sample. It holds

$$r_S = 1 - \frac{6}{n\left(n^2 - 1\right)} \sum_i \left(R_i - Q_i\right)^2 . \qquad (10)$$

Under the hypothesis of independence r_S has the mean value 0, the variance $\frac{1}{n-1}$ and approximately for n > 30 asymptotically normal distribution with critical values

$$r_S(n, \alpha) = \frac{u(\frac{\alpha}{2})}{\sqrt{n-1}} . \qquad (11)$$

If both variables are ordinal, we can use some nonparametric correlation coefficient, e.g. Kendall correlation coefficient. One option of the Kendall correlation coefficient is Goodman-Kruskal coefficient γ. It is evaluated from the count of the concordances *P* and discordances *Q*. It is appropriate coefficient to describe association of ordinal variables in the contingency table. Coefficient γ can be evaluated from the formula

$$\gamma = \frac{P - Q}{P + Q} . \qquad (12)$$

Further option of Kendall correlation coefficient is Kendall τ_c coefficient

$$\tau_c = \frac{2m(P-Q)}{n^2(m-1)}, \tag{13}$$

m is smaller dimension from the contingency table. It is appropriate coefficient to describe association in the table with various values of dimensions. Other option of Kendall correlation coefficient is τ_b coefficient

$$\tau_b = \frac{P-Q}{\sqrt{(P+Q+T_X)(P+Q+T_Y)}}, \tag{14}$$

T_x is the count of the pairs with the same value of the variable *X* and different value of the variable *Y*, T_Y is the count of the pairs with the same value of the variable *Y* and different value of the variable *X*. From the difference $(P-Q)$ we can calculate also Sommers correlation coefficients which are appropriate to evaluate if row or column variable can predict values of column or row variable.

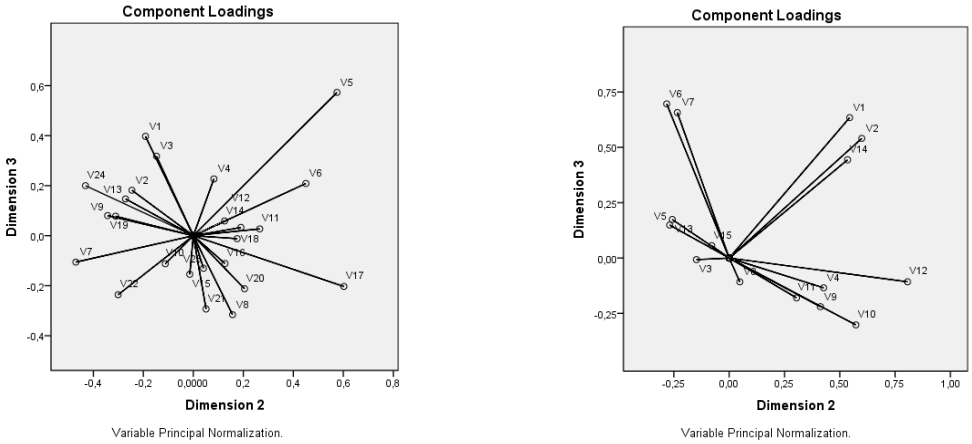
3 Results

Two datasets were evaluated by statistical software SPSS and R. One dataset is related to the research „Public perception of the policeman“. Questionnaire survey was held in the years 1995, 1999, 2006 in the Police academy of the Czech Republic and investigated typical policeman perception in the public society. Data from 100 respondents from the last realization (2006) are used in this article. Questionnaire survey included 24 questions (variables) with bipolar scales from 1 to 7, 4 is the neutral level. Lower positions mean positive, upper negative evaluation of the typical policeman. The questions described usually moral characteristics (adjectives) of the policeman, e.g. good-bad, active-passive, fast-slow etc. For more details see Moulisová 2009. The second dataset is related to the research „University students active lifestyle“ (15 selected Likert scaled variables from 100 respondents, scale 1-8), Students described satisfaction concerning different points of view of the students' life. Respondents evaluated their satisfaction on the scale from 1 (no satisfaction) to 8 (very satisfied). For more details see Valjent 2010.

Compared methods are categorical principal component analysis (procedure CATPCA in SPSS), multidimensional scaling (procedure ALSCAL and PROXSCAL in SPSS) and latent class models (procedure poLCA in R). The number of latent variables was

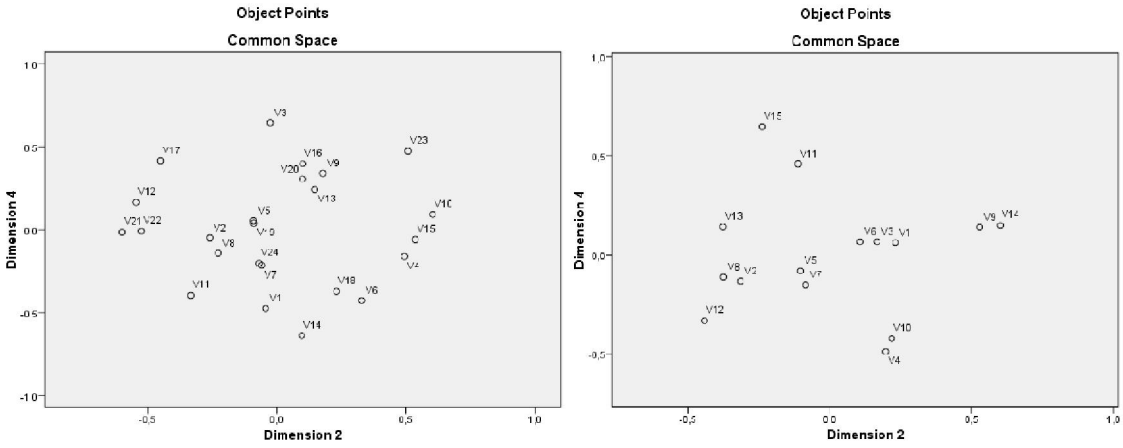
calculated with respect to the number of eigenvalues distinctly higher than one from categorical principal component analysis. It means 4 latent variables for the first dataset and 5 latent variables for the second dataset. For the comparison the same number of latent variables or classes was selected also for remaining dimensionality reduction methods. Kendall rank correlation coefficient we used as an inter-object distance measure and finally Spearman rank correlation coefficient between distances in original and reduced space was evaluated, see Tab. 1. Graphs of component loadings (Fig. 1) and coordinates from multidimensional scaling (Fig. 2, 3) are attached as well.

Fig. 1: Component loadings (CATPCA in SPSS, datasets 1, 2)



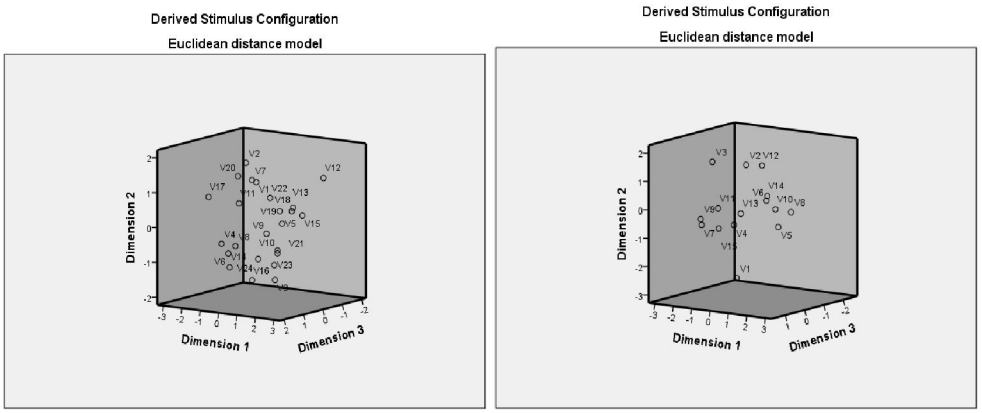
Source: own research

Fig. 2: Coordinates (MDS Proxscal in SPSS, datasets 1, 2)



Source: own research

Fig. 3: Coordinates (MDS Alscal in SPSS, datasets 1, 2)



Source: own research

Tab. 2: Spearmann rank corellation coefficients of inter-object distances in original and reduced space

Method	Dataset 1	Dataset 2
CATPCA	0.33	0.51
MDS PROXSCAL	0.18	0.25
MDS ALSCAL	0.20	0.28
LCA	0.29	0.48

Source: own research

Conclusion

For the comparison of dimensionality reduction methods applied to ordinal datasets we used Spearmann rank correlation coefficient between distances in original and reduced space. From the results of four dimensionality reduction methods applied to two ordinal dataset we can see, that satisfactory goodness of the data structure was obtained in case of CATPCA and LCA, weaker results were reached from the methods of multidimensional scaling. In further research other comparison techniques will be provided, e.g. Procrustes analysis and the results from all these techniques will be compared.

References

1. Hebák, P. et al. (2007). *Vícerozměrné statistické metody 3*. Praha: Informatorium
2. Hendl, J. (2006). *Přehled statistických metod: analýza a metaanalýza dat*. Praha: Portál

3. Holland, S. (2008). *Non-metric multidimensional scaling* (*mds*). Athens: R forge
4. Le, S., Josse, J., & Husson, F. (2008). Factominer: An R package for multivariate analysis. *Journal of statistical software*, 25(1)
5. Li, S., Vel, O., & Coomans, D. (1995). *Comparative performance analysis of non-linear dimensionality reduction methods*. North Australia: James Cook University
6. Linzer, D., & Lewis, J. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software*, 42(10)
7. Maaten, L. P. J., Postma, L. O., & Herik, H. J. (2008). *Dimensionality reduction: a comparative review*. Elsevier.
8. Moulisová, M. (2009). Výzkum percepce policisty. *Kriminalistika*, 42(1), 56-71.
9. Sobíšek, L., & Řezanková, H. (2011). Srovnání metod pro redukci dimenzionality aplikovaných na ordinální proměnné. *Acta Oeconomica Pragensia* , 1, 3-19
10. Valjent, Z. (2010). *Aktivní životní styl vysokoškoláků*. Praha: FEL ČVUT.

Contact

Martin Prokop

College of Polytechnics Jihlava

Tolstého 16

586 01 Jihlava

University of Economics, Prague

W. Churchill Square 4

130 67 Praha 3

maris@post.cz

Hana Řezanková

University of Economics, Prague

W. Churchill Square 4

130 67 Praha 3

hana.rezankova@vse.cz