DATA DIMENSIONALITY REDUCTION METHODS FOR ORDINAL DATA

Martin Prokop – Hana Řezanková

Abstract

From questionnaire survey we frequently get data, their values are expressed in ordinal (e.g. Likert) scale. The questionnaire contains usually a lot of questions, so we get multidimensional data matrix. To simplify calculations with the data it is useful to reduce dimensionality of the dataset. For ordinal data we use different or improved methods compared to quantitative data. This article includes the overview and comparison of dimensionality reduction methods (e.g. principal component analysis, factor analysis, multidimensional scaling, cluster analysis...). From these methods we get groups of similar variables (latent classes), in some cases we can create interpretation of these new variables.

Key words: ordinal data, dimensionality reduction, latent class models, multidimensional scaling, cluster analysis

JEL Code: C3, C6, C8

Introduction

Aim of this theoretical study is to describe data dimensionality reduction methods especially for ordinal data. This kind of data we frequently get from questionnaire surveys. Thus we solve the methods to reduce the number of variables characterizing individual objects. From big amount of variables (questions) we make new latent variables, which are created by groups of original variables. Conventional data dimensionality reduction methods usually assume quantitative variables, so we have to use modified or different statistical methods. There exist several methods sometimes with different results, so we compare the results from various methods. Application of the methods in this text will be usually described in the software R. Some methods especially for categorical data are described in detail (latent class models).

1 Overview of 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, e.g. using rank corellation coefficient. For more details see Hebák (2007).

Common methods of latent variables identification are latent class models. There exist many methods and different methods are available in statistical software packages, e.g. latent class cluster models LCC, discrete factor analysis models DFactor, latent trait analysis LTA, latent profile analysis LPA, latent class regression models LCR etc. For some of these methods in detail see Sobíšek and Řezanková (2011).

2 Principal component analysis

Some methods are based on multidimensional space projection into the space with lower dimension. Basic method is principal component analysis. The aim is to find real dimension of the data. To find 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 rank s. These two vectors are related by the transition formula, e. g. in the case of PCA (equations 1 and 2) 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 individual i on the axis s, G_s denotes the coordinate of the variable k on the axis s, λ_s the eigenvalue associated with the axis *s*, m_k the weight associated to the variable k, p_i the weight associated to the individual *i*.

Suitable count of components goes from the variance, which is explained by the sum of the variance of original variables or from screeplot of eigenvalues or from the count of eigenvalues, which are higher than 1, if we use correlation matrix instead of covariance matrix. Suitability and limitation of this method consists in the result, when we get from high count of variables small count of components with high proportion of explained variability.

High dependence of observed variables is also suitable, strong correlation among original variables and components too.

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. According to Sebastien et al. (2008) although a PCA applied on categorical data would yield results comparable to those obtained from a Multiple Correspondence Analysis (factor scores and eigenvalues are linearly related), there are more appropriate techniques to deal with mixed data types, namely Multiple Factor Analysis for mixed data available in the FactoMineR R package. Multiple Factor Analysis from the same package is also an option.

3 Multidimensional scaling

Other method based on multidimensional space projection into the space with lower dimension is multidimensional scaling MDS. Setting of axes (dimensions) is similar to PCA components setting. Multidimensional scaling is more general than factor analysis, because it is based on any relation matrix among variables or individuals. Method MDS is similar to cluster analysis, because it uses distance matrix of variables or individuals pairs. This distance can be based on similarity measure. Similarity of two variables can be estimated by some of mutual symmetric similarity measures. 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).

According to Holland (2008) nonmetric multidimensional scaling (MDS, also NMDS and NMS) is an ordination technique, that differs in several ways from nearly all other ordination methods. In most ordination methods, many axes are calculated, but only a few are viewed, owing to graphical limitations. In MDS, a small number of axes are explicitly chosen prior to the analysis and the data are fitted to those dimensions; there are no hidden axes of variation. Second, most other ordination methods are analytical and therefore result in a single unique solution to a set of data. In contrast, MDS is a numerical technique, that iteratively seeks a solution and stops computation when an acceptable solution has been found, or it stops after some pre-specified number of attempts. As a result, an MDS ordination is not a unique solution and a subsequent MDS analysis on the same set of data and following the same methodology will likely result in a somewhat different ordination. Third, MDS is not an eigenvalue-eigenvector technique like principal components analysis or correspondence analysis, that ordinates the data such that axis 1 explains the greatest amount of variance, axis 2 explains the next greatest amount of variance, and so on. As a result, an MDS ordination can be rotated, inverted, or centered to any desired configuration.

Unlike other ordination methods, MDS makes few assumptions about the nature of the data. For example, principal components analysis assumes linear relationships and reciprocal averaging assumes modal relationships. MDS makes neither of these assumptions, so is well suited for a wide variety of data. MDS also allows the use of any distance measure of the samples, unlike other methods, which specify particular measures, such as covariance or correlation in PCA or the implied chi-squared measure in detrended correspondence analysis.

The method starts with a matrix of data consisting of *n* rows of samples and *p* columns of variables, From this symmetrical matrix of all pairwise distances among samples 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 samples in starting configuration are calculated, typically with a Euclidean metric. These distances are regressed against the original distance matrix and the predicted ordination distances for each pair of samples 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. from equation 3 with one of the most common being Kruskal's Stress

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

where d_{hi} is the ordinated distance between samples *h* and *i*, and $\hat{d}$ is the distance predicted from the regression. This configuration is then improved by moving the positions of samples in ordination space by a small amount in the direction of steepest descent, the direction in which stress changes most rapidly. The ordination distance matrix is recalculated, the regression performed again and stress recalculated, and this entire procedure of nudging samples and recalculating stress is repeated until some small specified tolerance value is achieved or until the procedure converges by failing to achieve any lower values of stress, which indicates that a minimum (perhaps local) has been found.

A scree diagram (stress versus number of dimensions) can then be plotted, on which one can identify the point beyond which additional dimensions do not substantially lower the stress value. A second criterion for the appropriate number of dimensions is the interpretability of the ordination, that is, whether the results make sense. Stress increases both with the number of samples and with the number of variables.

R has two main MDS functions available, isoMDS, which is part of the MASS library, and metaMDS, which is part of the vegan library. The metaMDS routine allows greater automation of the ordination process, so is usually the preferred method. The metaMDS function uses isoMDS in its calculations as well as several helper functions. The metaMDS routine also has the useful default behavior of following the ordination with a rotation via principal components analysis such that MDS axis 1 reflects the principal source of variation and so on, as is characteristic of eigenvalue methods.

3 Overview of latent class and latent class regression models

According to Linzer (2011) latent class analysis is a statistical technique for the analysis of multivariate categorical data. When observed data take the form of a series of categorical responses (as for example, in public opinion surveys, individual-level voting data, studies of inter-rater reliability, or consumer behavior and decision-making), it is often our interest to investigate sources of confounding between the observed variables, identify and characterize clusters of similar cases, and approximate the distribution of observations across many variables of interest. Latent class models are a useful tool for accomplishing these goals. The latent class model seeks to stratify the cross-classiffcation table of observed (manifest) variables by an unobserved (latent) categorical variable, that eliminates all confounding between the manifest variables. Responses to all of the manifest variables are assumed to be statistically independent. The model, in effect, probabilistically groups each observation into a latent class, which in turn produces expectations about how that observation will respond on each manifest variable. Although the model does not automatically determine the number of latent classes in a given data set, it does offer a variety of parsimony and goodness of fit statistics, that the we may use in order to make a theoretically and empirically assessment.

Because the unobserved latent variable is nominal (membership of a class), the latent class model is actually a type of finite mixture model. The component distributions in the mixture are cross-classiffcation tables of equal dimension to the observed table of manifest variables, and, following the assumption of conditional independence, the frequency in each cell of each component table is simply the product of the respective class-conditional marginal frequencies (the parameters estimated by the latent class model are the proportion of observations in each latent class, and the probabilities of observing each response to each manifest variable, conditional on latent class). A weighted sum of these component tables forms an approximation (or, density estimate) of the distribution of cases across the cells of the observed table. Observations with similar sets of responses on the manifest variables will tend to cluster within the same latent classes. An extension of this basic model permits the inclusion of covariates to predict latent class membership. Whereas in the basic model, every observation has the same probability of belonging to each latent class prior to observing the responses to the manifest variables, in the more general latent class regression model, these prior probabilities vary by individual as a function of some set of independent variables.

poLCA is a software package for the estimation of latent class and latent class regression models for polytomous outcome variables (variables with more than two distinct categories), implemented in the R 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. poLCA uses expectation-maximization and Newton-Raphson algorithms to find maximum likelihood estimates of the model parameters.

4 Latent class models

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. Suppose we observe J polytomous categorical variables (the manifest variables), each of which contains K_j possible outcomes, for individuals $i = 1,...,N$. The manifest variables may have different numbers of outcomes, hence the indexing by j. 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 observation 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 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 individual will belong to each class before taking into account the responses Y_{ijk} provided on the manifest variables. The probability that an individual i in class r produces a particular set of J outcomes on the manifest variables, assuming conditional independence of the outcomes Y given class memberships, is the product

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

The probability density function across all classes is the weighted sum

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

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 individual belongs to each class, conditional on the observed values of the manifest variables, can be calculated using Bayes' formula:

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

where $r_i \in \{1,...R\}$. Recall that the $\hat{\pi}_{jrk}$ are estimates of outcome probabilities conditional on class r. 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 observations, 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. poLCA estimates the latent class model by maximizing the log-likelihood function

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

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 individual's class membership is unknown and may be treated as missing data. The EM algorithm proceeds iteratively. Begin 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, calculate the missing class membership probabilities using Equation 6, substituting in $\hat{p}_r^{old}$ and $\hat{\pi}_{jrk}^{old}$. In the maximization step, update the parameter estimates by maximizing the log-likelihood function given these posterior $\hat{P}(r_i \mid Y_i)$ with

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

as the new prior probabilities and

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

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.

poLCA takes advantage of the iterative nature of the EM algorithm to make it possible to estimate the latent class model even when some of the observations on the manifest variables are missing. Although poLCA does offer the option to listwise delete observations with missing values before estimating the model, it is not necessary to do so. Instead, when determining the product in Equation 4 and the sum in the numerator of Equation 9, poLCA simply excludes from the calculation any manifest variables with missing observations. The priors are updated in Equation 6 using as many or as few manifest variables as are observed for each individual. Depending on the initial values chosen for $\hat{p}_r^{old}$ and $\hat{\pi}_{jrk}^{old}$, and the complexity of the latent class model being estimated, the EM algorithm may only find a local maximum of the log-likelihood function, rather than the desired global maximum. For this reason, it is always advisable to re-estimate a particular model a couple of times when using poLCA, in an attempt to find the global maximizer to be taken as the maximum likelihood solution.

One of the benefits of latent class analysis, in contrast to other statistical techniques for clustered data, is the variety of tools available for assessing model fit and determining an appropriate number of latent classes R for a given data set. In some applications, the number of latent classes will be selected for primarily theoretical reasons. In other cases, however, the analysis may be of a more exploratory nature, with the objective being to locate the best fitting or most parsimonious model. We may then begin by fitting a complete independence model with $R = 1$, and then iteratively increasing the number of latent classes by one until a suitable fit has been achieved. Adding an additional class to a latent class model will increase the fit of the model, but at the risk of fitting to noise, and at the expense of estimating a further $1 + \sum_j (K_j - 1)$ model parameters. Parsimony criteria seek to strike a balance between over- and under-fitting the model to the data by penalizing the log-likelihood by a function of the number of parameters being estimated. The two most widely used parsimony measures are the Bayesian information criterion, or BIC and Akaike information criterion, or AIC. Preferred models are those that minimize values of the BIC and/or AIC. Let Λ represent the maximum log-likelihood of the model and Φ represent the total number of estimated parameters. Then, AIC = $-2\Lambda+2\Phi$ and BIC = $-2\Lambda+ \Phi \ln N$. poLCA calculates these parameters automatically when estimating the latent class model. The BIC will usually be more appropriate for basic latent class models because of their relative simplicity. Calculating Pearson's X^2 goodness of fit and likelihood ratio chi-square (G^2) statistics for the observed versus predicted cell counts is another method to help determine how well a particular model fits the data, for more details see Linzer (2011). Like the AIC and BIC, these statistics are outputted automatically after calling poLCA.

5 Optimal scaling by Gifi methods

The challenge with categorical variables is to find a suitable way to represent distances between variable categories and individuals in the factorial space. To overcome this problem, we can look for a non-linear transformation of each variable, whether it is nominal, ordinal, polynomial, or numerical with optimal scaling. This is well explained in Leeuw (2009), and an implementation is available in the corresponding R package homals. As extension to this transformation, having nonmetric variables, we can use dimensionalty reduction methods, e.g. nonlinear principal component analysis (NLPCA). The term nonlinear pertains to nonlinear transformations of the observed variables. In Gifi terminology, NLPCA can be defined as homogeneity analysis with restrictions on the quantification matrix.

6 Fuzzy Clustering

According to Oksanen (2010) we have so far worked with classiffcation methods, which implicitly assume, that there are distinct classes. The real situation is usually different. If there are classes, they are vague and have intermediate and untypical cases. With one word, they are fuzzy. Fuzzy classiffcation means, that each observation has a certain probability of belonging to a certain class. In the crisp case, it has probability 1 of belonging to a certain class, and probability 0 of belonging to any other class. In a fuzzy case, it has probability < 1 for the best class, and probabilities > 0 for several other classes. Fuzzy classiffcation is similar to K-means clustering in finding the optimal classiffcation for a given number of classes, but the produced classiffcation is fuzzy: the result is a probability profile of class membership. The fuzzy clustering is provided by function fanny (equation 10) in package cluster. Requested membership probabilities we get from the minimalization of the function

$$J_F = \sum_{h=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ih}^2 u_{jh}^2 d_{ij}}{2\sum_{j=1}^{n} u_{ih}^2} \tag{10}$$

where the values of u are membership probabilities and values of d are Euclidean distances among the objects. It is difficult to show the fuzzy results graphically, but it is possible to use stars function (with many optional parameters) to show the probability profile, and it draws a convex hull of the crisp classification. The size of the sector shows the probability of the class membership and in clear cases one of the segments is dominant.

Conclusion

This study was aimed to the overview of data dimensionality reduction method especially for categorical (ordinal) data. Some advantages and difficulties of the methods were presented, in future research these methods will be apllied on real dataset and comparison of the results will be performed.

References

1. Hebák, Petr. *Vícerozměrné statistické metody 3*. Praha: Informatorium, 2007.
2. Hendl, Jan. *Přehled statistických metod: analýza a metaanalýza dat*. Praha: Portál, 2006.
3. Holland, Steven M. *Non-metric multidimensional scaling (MDS)*. Athens: R forge, 2008.
4. Le, Sebastien, and Josse, Julie, and Husson, Francois. “FactoMineR: An R package for multivariate analysis.“ *Journal of statistical software* 25 March 2008: sec. 1.

5. Leeuw, Jan, and Mair, Patrick. “Gifi Methods for Optimal Scaling in R: The Package homals.“ *Journal of statistical software* 31 Aug. 2009: sec. 4.
6. Linzer, Drew, and Lewis, Jeffrey. “poLCA: An R Package for Polytomous Variable Latent Class Analysis.“ *Journal of statistical software* 42 June 2011: sec. 11.
7. Oksanen, Jari. *Cluster Analysis: Tutorial with R*. Mendeley, 2010.
8. Sobíšek, Lukáš, and Řezanková, Hana. “Srovnání metod pro redukci dimenzionality aplikovaných na ordinální proměnné.“ *Acta Oeconomica Pragensia* 1. 2011: 3-19.

Contact

Martin Prokop
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4
130 67 Praha 3

maris@post.cz

Hana Řezanková
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4
130 67 Praha 3

hana.rezankova@vse.cz