Courant Research Centre

‘Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis’

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 172

Treatments of Non-metric Variables in Partial Least Squares and Principal Component Analysis

Jisu Yoon, Tatyana Krivobokova

March 2015

Wilhelm-Weber-Str. 2 · 37073 Goettingen · Germany
Phone: +49-(0)551-3914066 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

Treatments of Non-metric Variables in Partial Least Squares and Principal Component Analysis

Jisu Yoon[1] Tatyana Krivobokova [1]

March 24, 2015

Abstract

This paper reviews various treatments of non-metric variables in Partial Least Squares (PLS) and Principal Component Analysis (PCA) algorithms. The performance of different treatments is compared in the extensive simulation study under several typical data generating processes and recommendations are made. An application of PLS and PCA algorithms with non-metric variables to the generation of a wealth index is considered.

[1] Courant Research Center "Poverty, Equity and Growth", Georg-August-Universität Göttingen, Wilhelm-Weber-Str. 2, 37073 Göttingen, Germany, E-mail: jisu.yoon@zentr.uni-goettingen.de

1 Introduction

Principal Component Analysis (PCA, Hotelling, 1933) and Partial Least Squares (PLS, Wold, 1966) are popular dimension reduction techniques, which are typically applied in case of multicollinear predictors and are also often used to build various composite indices. Both PCA and PLS are developed for the analyses of metric variables. However, in practice one often is faced with non-metric variables. Even though there is a large number of approaches to treat non-metric variabels in PCA and PLS algorithms available in the literature, it is not always clear under which assumptions about the data generating process (DGP) these algorithms perform best. To the best of our knowledge, there is no clear guideline for practitioners how to select the best treatment of non-metric variables for data at hand. In this work we review various treatments of non-metric variables for PCA and PLS algorithms. All together, we consider eleven methods grouped into three main types. All treatments for non-metric variables are described in detail, together with necessary assumptions, if appropriate. An extensive simulation study aims to compare the performance of all methods under several typical data generating processes and to make recommendations for practitioners.

As an application, we consider construction of a wealth index with PCA and PLS. Wealth indices (Filmer and Pritchett, 2001; Rutstein and Johnson, 2004) are composite indices that aim to measure household wealth based on the posession of certain assets. In general, a composite index is an aggregated variable comprising individual indicators and weights that commonly represent the relative importance of each indicator (Nardo et al., 2005). Other examples of such indices include the KOF index of Globalization (Dreher, 2006) that quantifies globalization and the Social Institutions and Gender Index (SIGI; Branisa et al., 2013) that measures social institutional aspects of gender inequality across countries. The most crucial step in building an index is to determine appropriate weights, which is typically done with PCA or PLS. Since in practice many variables that enter such indices are non-metric, it is of great importance to apply appropriate methods for treating non-metric variables for PCA and PLS. Our wealth index application illustrates the generation and use of a composite index with non-metric variables. A wealth index is often used as a proxy for household expenditures, so that it is important to quantify how well the wealth index is able to predict household expenditures. Therefore, we perform regression analyses, where household expenditures are explained by the wealth index and a set of control variables. We perform a model selection with respect to the treatment of non-metric variables and the set of control variables to improve estimated prediction performance.

The rest of the paper is organized as follows. Section 2 recapitulates PCA and PLS algorithms and reviews the treatments of non-metric variables in PCA and PLS in the literature. In Section 3 the simulation study is presented, various treatments are compared and recommendations under several typical DGPs are made. The analysis on the wealth index is performed in Section 4, before we conclude in Section 5.

2 PCA and PLS with Non-metric variables

2.1 PCA and PLS Algorithms

First, we give a brief discription of standard PLS and PCA algorithms with metric variables. Let us consider the following regression model $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^N$ is a regressand vector and $X \in \mathbb{R}^{N \times K}$, $K < N$ is a regressor matrix. Both y and X are assumed to be centered. Regression coefficients are denoted by $\beta \in \mathbb{R}^K$ and $\varepsilon \in \mathbb{R}^N$ is the error term, such that $\mathbb{E}(\varepsilon|X) = 0$ and $\mathrm{Cov}(\varepsilon|X) = \sigma^2 I_n$.

PCA and PLS scores are built as linear combinations of regressors, that is $T = XW$, where $T = (t_1, ..., t_A) \in \mathbb{R}^{N \times A}$ is the score matrix and $W = (w_1, ..., w_A) \in \mathbb{R}^{K \times A}$ is the weight matrix with $A \leq K$. Thereby, the weight matrices are different in PCA and PLS.

PCA weights w_a are found from

$$w_a = \underset{\|\omega\|=1}{\operatorname{argmax}} \, \omega^T X^T X \omega, \text{ subject to } w_a \perp ... \perp w_1, \; a = 1, ..., A,$$

which is the a-th eigenvector of $X^T X$. The first PLS weight vector w_1 is given by

$$w_1 = \underset{\|\omega\|=1}{\operatorname{argmax}} \, (\omega^T X^T y)^2 = \frac{X^T y}{\|X^T y\|},$$

while the later weights w_a are found solving the same problem subject to the mutual orthogonality $w_a \perp ... \perp w_1$. We refer to de Jong (1993) for more details.

2.2 Treatments of Non-metric Variables in PCA and PLS

Treatments of non-metric variables in PCA ans PLS algorithms available in the literature can be organized into three main categories. The first group of methods uses certain transformations of each unique category of a non-metric variable into a variable. The second group of approaches applies various scalings of non-metric variables after which these variables are treated as metric. The last group of treatments assumes a certain continuous latent variable behind the observed non-metric variable and uses the variance-covariance matrix of the latent variables to calculate PLS or PCA weights. In the following a brief summary of these methods is given. Thereby, it is assumed that the first K_n columns of regressor matrix X contain non-metric variables, the j-th non-metric variable has m_j unique values, which are integers $x_{ij} \in \{0, 1, ..., m_j - 1\}$, $i = 1, \ldots, N$, $j = 1, \ldots, K$ and the regressand y is always metric.

First, consider methods which transform each unique category of a non-metric variable into a variable. These are **dummy coding** (Filmer and Pritchett, 2001), the **aggregation method** (Saisana and Tarantola, 2002), **regular simplex method** (Niitsuma and Okada, 2005) and **multiple correspondence analysis** (**MCA**; Greenacre, 2010). All those methods require no particular distributional assumptions on variables in X. **Dummy coding** transforms each unique value of a non-metric variable to a dummy variable. In other words, one replaces x_{ij} with $\tilde{x}_{ij} = (I(x_{ij} = 0), I(x_{ij} = 1), ..., I(x_{ij} = m_j - 1)) \in \mathbb{R}^{1 \times m_j}$, where I denotes the indicator function. The first element may be dropped for an easier interpretation. The **aggregation method** in this paper is defined as a cluster level average. That is, it is assumed that each observation x_{ij} belongs to a cluster $c \in \{1, ..., C\}$ and it is replaced with $\tilde{x}_{ij} = (A_{c,j}(0), A_{c,j}(1), ..., A_{c,j}(m_j - 1)) \in \mathbb{R}^{1 \times m_j}$, where $A_{c,j}(u) = \left(\sum_{i \in c} I(x_{ij} = u)\right) \left(\sum_{i \in c} \sum_{v=0}^{m_j - 1} I(x_{ij} = v)\right)^{-1}$. The **regular simplex method** transforms each value of a non-metric variable to a corresponding vertex coordinate of a regular simplex, that is $\tilde{x}_{ij} = \text{Ver}_{m_j - 1}(x_{ij}) \in \mathbb{R}^{1 \times m_j}$, where $\text{Ver}_{m_j - 1}(x_{ij})$ transforms x_{ij} to the $(x_{ij} + 1)$-th vertex coordinate in $m_j - 1$ dimension. For all three afore-mentioned methods non-metric variables after the treatment and metric variables are concatenated, resulting in a row $\tilde{X}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, ..., \tilde{x}_{iK_n}, x_{iK_n+1}, ..., x_{iK})$ of matrix $\tilde{X}$. Finally, usual PLS or PCA is applied on $\tilde{X}$. The last approach in this group, **MCA**, first discretizes metric variables, so that the regressor matrix contains only non-metric variables. Afterwards, the regressor matrix is transformed to an indicator matrix using dummy coding without dropping the first column, which will be denoted by Z. Subsequently, Z is standardized as $Z_s = \text{diag}(r^{-1/2})(P - rc^T)\text{diag}(c^{-1/2})$, where $P = Z(\underline{1}^T Z \underline{1})^{-1}$, $r = P\underline{1}$, $c = P^T \underline{1}$ and $\underline{1}$ denotes a vector of 1s of the appropriate length. Finally, Singular Vector Decomposition (SVD) is applied to Z_s and the left singular vectors are used as scores. This procedure can be interpreted as a PCA on discretized regressors with a special dummy coding, where each column is weighted, so that categories with many incidences are equally important as categories with fewer incidences.

Second group of approaches applies certain scaling to each unique value of non-metric variables. These methods include the **optimal scaling method** (Tenenhaus and Young, 1985), **non-metric partial least squares regression** (**NM-PLSR**; Russolillo, 2009) and **categorical principal component analysis** (**CATPCA**; Meulman, 2000). No distributional assumptions on X are necessary. The **optimal scaling method** maximizes the sum of variances of non-metric variables in terms of the scaling of unique categories. First, an indicator matrix from non-metric variables Z is built and the eigenvector ν, corresponding to the second largest eigenvalue of $K^{-1}\text{diag}(\underline{1}^T Z)^{-1} Z^T Z$, is determined. Finally, PCA or PLS is applied to $\tilde{X} = (Z_1\nu_1, ..., Z_{K_n}\nu_{K_n}, x_{K_n+1}, ..., x_K)$, where $Z_j \in \mathbb{R}^{N \times m_j}$ and $\nu_j \in \mathbb{R}^{m_j}$ denote the columns of Z and the components of ν corresponding to variable j, $j = 1, \ldots, K_n$. Next approach, **NM-PLSR**, maximizes the covariance between the first score and regressand in term of the scaling of unique categories. The quantification function is defined as $Q(x_j, y) = Z_j(Z_j^T Z_j)^{-1} Z_j^T y / \left\| Z_j(Z_j^T Z_j)^{-1} Z_j^T y \right\|$, if x_j is treated as nominal. The quantification function for ordinal x_j is analogous, except that it is constrainted to respect the order. If the quantification of a category does not respect the order, another quantification is calculated after the category is merged to an adjacent category. Now PLS is run with $\tilde{X} = (\tilde{x}_1, ..., \tilde{x}_{K_n}, x_{K_n+1}, ..., x_K)$, where $\tilde{x}_j = Q(x_j, y)$, $j = 1, ..., K_n$. The quantification does not change for the later scores. The last method in this group, **CATPCA**, maximizes the sum of the variances of scores in terms of the scaling of unique categories. CATPCA allows to select the number of scores to be considered in the maximization, but in analogy to NM-PLSR, we opted for the case with only one score considered during the quantification. In our simulation studies and application CATPCA showed rather inferior performance. Therefore, we omit the details of this lengthy algorithm and refer to IBM SPSS Statistics (2013) for more details.

Polychoric PCA (Kolenikov and Angeles, 2009) is based on the assumption that observed ordinal variables are generated from a latent multivariate normal process discretized at some thresholds. Under this assumption, thresholds and variance-covariance matrix are estimated and PCA is performed on centered and autoscaled regressors using the eigenvectors from the variance-covariance matrix as the weights. In the following Φ and Φ_2 denote standard normal and bivariate standard normal cumulative distribution function, respectively, and ϕ is standard normal density function. First, one estimates the thresholds at which the latent normal variable is discretized. Let $\alpha_j = (\alpha_{j(-1)}, \alpha_{j0}, ..., \alpha_{jm_j-1}) \in \mathbb{R}^{m_j+1}$ be a vector of thresholds for variable x_j, where $\alpha_{ju} = \Phi^{-1}\left(N^{-1}(-0.5 + \sum_{i=1}^{N} I(x_{ij} \leq u))\right)$ for $u = 0, ..., m_j - 2$ and $\alpha_{j(-1)} = -\infty$, $\alpha_{jm_j-1} = \infty$. Second, the correlation between variables is estimated by maximizing likelihood conditional on the thresholds, i.e., $\rho = cor(\mathbf{X}_j, \mathbf{X}_{j'})$ and $\hat{\rho} = \underset{\rho}{\operatorname{argmax}} \ell(\rho)$, where $\ell(\rho) = \sum_{i=1}^{N} ln(L(x_{ij}, x_{ij'}|\rho, \alpha, \alpha^{'})))$. If one estimates the correlation between two ordinal variables, i.e., polychoric correlation, the likelihood for observation i is $L(x_{ij}, x_{ij'}|\rho, \alpha, \alpha^{'}) = \Phi_2(\alpha_{jx_{ij}}, \alpha_{j'x_{ij'}}|\rho) - \Phi_2(\alpha_{jx_{ij}-1}, \alpha_{j'x_{ij'}}|\rho) - \Phi_2(\alpha_{jx_{ij}}, \alpha_{j'x_{ij'}-1}|\rho) + \Phi_2(\alpha_{jx_{ij}-1}, \alpha_{j'x_{ij'}-1}|\rho)$. The correlation between a metric variable and an ordinal variable is called polyserial correlation. The likelihood for an observation with ordinal variable x_{ij} and metric variable $x_{ij'}$ is $L(x_{ij}, x_{ij'}|\rho, \alpha) = (\Phi(\alpha_{jx_{ij}} - \rho x_{ij'}) - \Phi(\alpha_{jx_{ij}-1} - \rho x_{ij'}))\phi(x_{ij'})$. We adapt polychoric PCA in the the PLS context, which we call **polyserial PLS**. This method applies autoscaling to regressors and outcome variable and finds the first PLS weights, $w_1 = \mathrm{Cor}(y, X)/\left\|\mathrm{Cor}(y, X)\right\|$, where $\mathrm{Cor}(y, X)$ is polyserial or Pearson correlation depending on whether regressor is ordinal or numerical. Kolenikov and Angeles (2009) discuss also the **normal mean coding**, which is a scaling approach based on the same distributional assumption as polychoric PCA. It scales each unique category of an ordinal variable to the expected value of the latent normal variable of the group, to which the category belongs. The scaling of x_{ij} is computed as $\mathbb{E}(x_{ij}^{*}|x_{ij}) = \int_{\alpha_{jx_{ij}-1}}^{\alpha_{jx_{ij}}} z\phi(z)dz = \phi(\alpha_{jx_{ij}-1}) - \phi(\alpha_{jx_{ij}})$, where x_{ij}^{*} denotes the underlying latent variable.

Additionally to the described three groups of methods, we study **ordinal PCA** and **ordinal PLS**, where ordinal variables are simply treated as if they were metric, see Kolenikov and Angeles (2009).

3 Simulations

In this section we describe the results of the simulation study that compares various treatments of non-metric variables for PCA and PLS algorithms under several data generating processes.

3.1 Simulation Design

We adapt the simulation designs from Naes and Martens (1985) and Kolenikov and Angeles (2009) with some adjustments. All simulation designs relay on a latent variable model (Muthén, 1984; Chin et al., 2003). A latent variable model explictly assumes latent variables, which are not directly observable, but manifested to other observable variables. For example, in a wealth index application, one cannot observe household wealth directly, but wealth is assumed to be manifested to household asset posessions, such as car, radio and bicycle, which are observable. A latent variable model reconstructs the latent concept based on the observed variables, which are manifested from the latent variable. To highlight the difference in PCA and PLS algorithms we design two main DGPs as follows. Under the first data generating process (**DGP 1**), covariates of the model contain only one latent factor, which is related to the response. In this setting both PCA and PLS algorithms are expected to perform similarly and the main focus is on various methods for non-metric variables. Under the second data generating process (**DGP 2**), covariates of the model contain two latent factors: the first one is related to the regressand and the second one is not. Thereby, the variance of the second latent factor, which is unrelated to the response variable, is much larger than that of the first latent factor. Hence, PLS algorithm, which maximizes the covariance between the response and covariates, remains unaffected by the unrelated latent factor with large variance and should perform much better than PCA, which maximizes the covariance of covariates and, hence, is highly influenced by the “spurious” covariates related to the second latent factor. In this setting we aim not only to demonstrate the performance of methods for non-metric variables, but also to compare PCA and PLS methods. DGP 1 has a practical relevance, when the largest variations in the observed variables come from the latent variable of interest, e.g., in a wealth index application, the posession of a car, house and so on could be largely determined by household wealth. DGP 2 is relevant to the case, where the observed variables include only small variations from the latent variable of interest, while the observed variables are influenced by other factors too. For example, one may try to measure globalization by the number of IKEA shops in a country. But the number of IKEA shops is not only determined by globalization, but also by local demand, competitors, regulations, etceteras, which may account for the main variations in the observed variable. Finally, **DGP 1H** and **DGP 2H** introduce heterogeneity of observations to DGP 1 and DGP 2. These settings reflect practical situations with different clusters in the data. For example, African countries show different behaviors than other countries in terms of economic growth (Barro, 1989; Sachs and Warner, 1997). When one studies a survey data such as Demographic and Health Surveys (Central Bureau of Statistics (CBS) Kenya et al., 2004), certain covariates may have different contributions for observations measured in urban and rural areas or male and females.

Formal definitions of all data generating process are as follows. **DGP 1** corresponds to the following model. Let

$$x_{ij}^* = \Xi_{i1}\lambda_{1j} + \Delta_{ij}, \;\; i = 1, \ldots, N, \;\; j = 1, \ldots, K.$$

Here $\lambda_{1j} = 1/\sqrt{K}$, $j = 1, \ldots, K$ are loadings and Ξ_{i1} is the common latent factor, which is distributed either as $\Xi_{i1} \sim \mathcal{N}(0, 1)$ or $\Xi_{i1} \sim \ln \mathcal{N}(-1.44, 1.55)$. The parameters of the log normal distribution imply variance 1 and skewness 13. Error terms $\Delta_i = (\Delta_{i1}, ..., \Delta_{iK})$ are the unique factors with $\Delta_i \sim \mathcal{N}_K\left(0_K, I_K/(9K)\right)$, such that the signal to noise ratio

$\sqrt{\sum_{j=1}^{K} \mathrm{Var}(\Xi_{i1}\lambda_{1j}) / \sum_{j=1}^{K} \mathrm{Var}(\Delta_{ij})} = 3$. Row vector $X_i^* = (x_{i1}^*, ..., x_{iK}^*)$ denotes the i-th observation in the regressor matrix and the superscript $*$ states that these are metric variables before discretization. The latent factor is connected to the outcome variable y_i as

$$y_i = \Xi_{i1}\beta_1 + \varepsilon_i, \ \ i = 1, \ldots, N, \tag{1}$$

where $\beta_1 = 1$ and the error term $\varepsilon_i \sim \mathcal{N}(0, 0.01)$. Hence, the only latent factor is connected to the outcome variable and in this setting one can expect both PCA and PLS to perform equally well.

DGP 2 introduces an additional factor with large variance which does not influence the response variable:

$$x_{ij}^* = \Xi_{i1}\lambda_{1j} + \Xi_{i2}\lambda_{2j} + \Delta_{ij},$$

where $(\Xi_{i1}, \Xi_{i2}) \sim \mathcal{N}_2\big(0_2, \left(\begin{smallmatrix} 1 & 0 \\ 0 & 5 \end{smallmatrix}\right)\big)$ or $(\Xi_{i1}, \Xi_{i2}) \sim \ln\mathcal{N}_2\big((-1.44, -0.63), \left(\begin{smallmatrix} 1.55 & 0 \\ 0 & 1.55 \end{smallmatrix}\right)\big)$, so that the parameters of the log normal distribution imply variances 1 and 5 for Ξ_{i1} and Ξ_{i2}, respectively, and skewness 13 for both. The loadings λ_{1j} are as before, while λ_{2j} are chosen so that $\|\lambda_1\| = \|\lambda_2\| = 1$ and $\lambda_1 \perp \lambda_2$. The distribution of $\Delta_i = (\Delta_{i1}, ..., \Delta_{iK})$ is the same as in DGP 1, but the signal to noise ratio increases to $3\sqrt{6}$. The model for the outcome variable remains unchanged, i.e., (1) still holds, so that Ξ_{i2} does not have any influence on y_i. In this setting PLS is expected to outperform PCA, since by defintion it remains unaffected by the second latent factor with large variance, in contrast to PCA.

DGP 1H and **DGP 2H** introduce a Boolean variable which interacts with the first latent factor of DGP 1 and 2, respectively, that is

$$y_i = \Xi_{i1}\beta_1 + D_i\beta_2 + \Xi_{i1} \circ D_i\beta_3 + \varepsilon_i,$$

with $D_i \sim \mathrm{Bin}(1, 0.5)$, $\beta_2 = \beta_3 = 1$ and $\circ$ denoting the Hadamard product. This is a simple example of heterogenous observations. In applications such heterogeneity appears, if the regression coefficients differ among different clusters. Neglecting such heterogenous observations should lead to a deterioration of the performance, which we would like to quantify in our simulation study and determine which methods stay robust.

In the next step, we discretize some variables in X^*. The discretization of the j-th variable x^*_{ij} with m_j number of unique categories is performed by the following function.

$$x_{ij} = \begin{cases} m_j - 1, & \text{if} \quad \tau_{j,m_j-1} < x^*_{ij} \\ m_j - 2, & \text{if} \quad \tau_{j,m_j-2} < x^*_{ij} \leq \tau_{j,m_j-1} \\ \vdots & \qquad \vdots \\ 1, & \text{if} \quad \tau_{j,1} < x^*_{ij} \leq \tau_{j,2} \\ 0, & \text{if} \quad x^*_{ij} \leq \tau_{j,1}, \end{cases}$$

where $\tau_j = (\tau_{j,1}, ..., \tau_{j,m_j-1})$ are some thresholds for x^*_{ij}. The thresholds are generated as $\tau_j = (\tau_{j,1}, ..., \tau_{j,m_j-1}) = (F^{-1}(u_{j,1}), ..., F^{-1}(u_{j,m_j-1}))$, where $F(\cdot)$ is the empirical CDF of the realizations of x^*_{ij} and $u_{j,1}, ..., u_{j,m_j-1}$ are generated from the uniform distribution on [0,1] and sorted ascending.

To measure the performance of various non-metric PCA and PLS methods, the mean squares error of prediction (MSEP) is calculated from a Monte Carlo sample of 500 repetitions. The MSEP in the l-th iteration is defined as

$$MSEP_l = \frac{1}{N}(\Xi_{1l}\beta_1 - U_l\hat{\gamma}_l)^T(\Xi_{1l}\beta_1 - U_l\hat{\gamma}_l)$$

for DGP 1 and 2 and for DGP 1H and 2H as

$$MSEP_l = \frac{1}{N}(\Xi_{1l}\beta_1 + D_l\beta_2 + \Xi_{1l} \circ D_l\beta_3 - U_l\hat{\gamma}_l)^T(\Xi_{1l}\beta_1 + D_l\beta_2 + \Xi_{1l} \circ D_l\beta_3 - U_l\hat{\gamma}_l),$$

where $\Xi_{1l} = (\Xi_{11l}, ..., \Xi_{N1l})$ and $D_l = (D_{1l}, ..., D_{Nl})$. The matrix $U_l = (\underline{1}, t_{1l}) \in \mathbb{R}^{N \times 2}$ includes the intercept with the first PLS or PCA score and $\hat{\gamma}_l$ is the OLS coefficients of $y_l = (y_{1l}, ..., y_{Nl})$ on U_l. True values $\Xi_{1l}\beta_1$ and $\Xi_{1l}\beta_1 + D_l\beta_2 + \Xi_{1l} \circ D_l\beta_3$ are scaled as unit variance before fitting the models to make the MSEPs from different settings comparable.

We consider the following settings under each DGP. The sample size N is either 100 or 1000 and the number of regressors K is either 10 or 50. The proportion of non-metric variables in the regressor matrix is 50% or 80%. The expected number of categories of non-metric variables m_j is either 3 or 7. Thereby m_j is generated from the Poisson distribution with mean $\lambda = 1$ or $\lambda = 5$ and we add 2 to m_j to guarantee at least two unique values in a variable.

PLS and PCA solutions are known to depend on the scaling of regressors (Wold et al., 2001; Keun et al., 2003). Scaling approaches, as well as polychoric PCA and polyserial PLS, by definition imply particular scalings of regressors. For dummy coding method we compare three scaling approaches: no scaling, autoscaling and block scaling. Auto-scaling centers and standardizes regressors to the unit variance, while block scaling sets the sum of the variances of dummy variables from one non-metric variable to one.

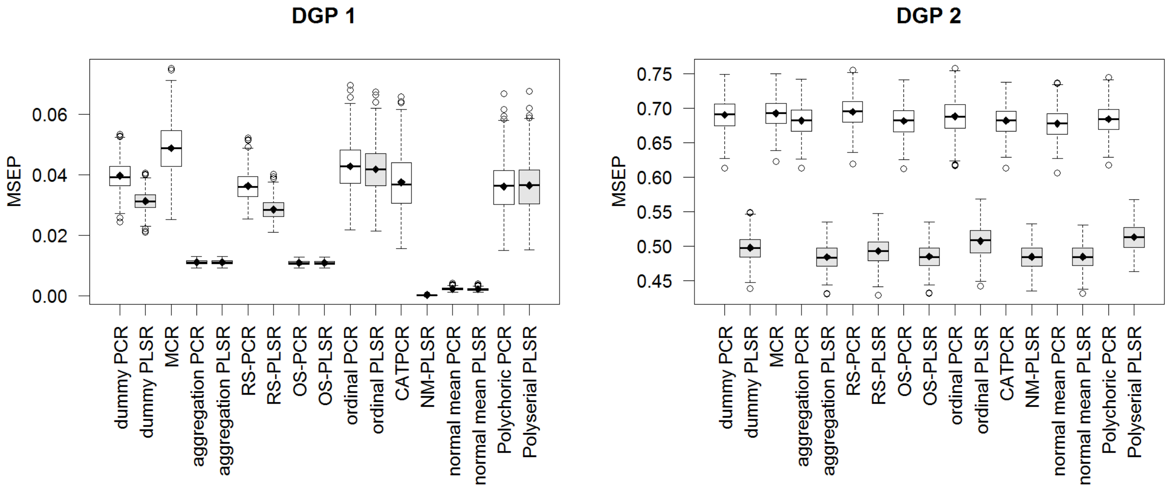
Note that our model is restricted to just one latent component and only the first PCA and PLS scores are estimated, implicitly assuming that the number of latent components is known. This allows us to exclude the variability due to the estimation of the number of latent components, so that the comparison beween the methods is not influenced by an extra variability. Moreover, in many applications only the first PCA or PLS components is of interest and is estimated.

3.2 Simulation Results

The simulation results are reported via box plots, where means are marked with black dots. We define *Base setting 1* as DGP 1, normally distributed Ξ_1, $N = 1000$, $K = 50$, proportion of non-metric variables is 80% and expected number of categories is 7. *Base setting 2* is the same as *Base setting 1*, except that DGP 2 is used instead of DGP 1.

The reported methods in the box plots are PCA or PLS with dummy coding (dummy PCR/PLSR), the aggregation method (aggregation PCR/PLSR), the regular simplex method (RS-PCR/PLSR), the optimal scaling method (OS-PCR/PLSR), the ordinal PCR/PLSR, the normal mean coding (normal mean PCR/PLSR), MCA (MCR), NM-PLSR, CATPCR, polychoric PCR and polyserial PLSR. For dummy coding only the results with no scaling are reported, because other scaling approaches perform similar or worse for the selected settings. For similar reasons, both NM-PLSR and CATPCR with only nominal quantification are reported.

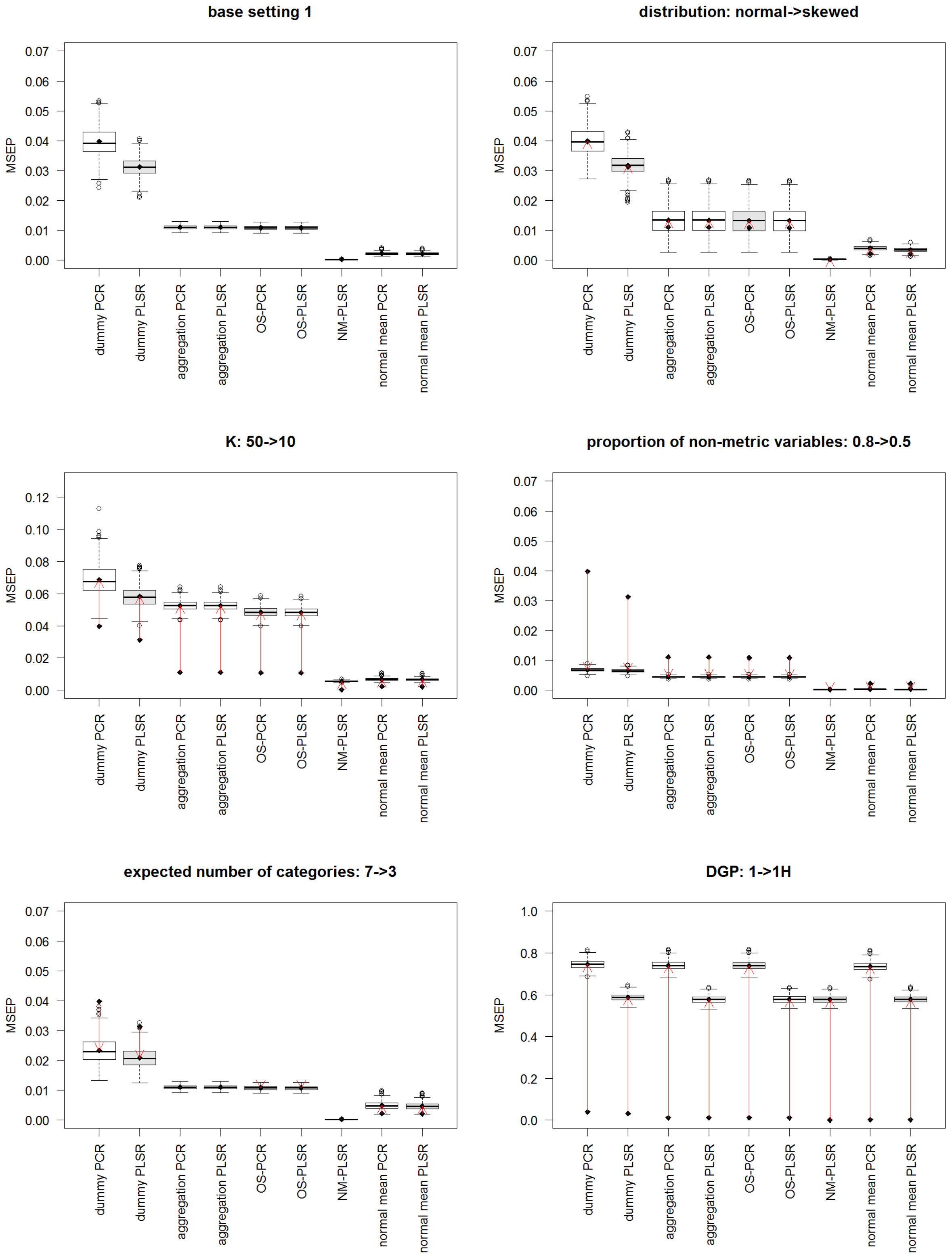
Figure 1: MSEP under DGP 1 (left) and DGP 2 (right)



Base setting 1 and 2 are reported. PCA-based methods are colored white and PLS-based methods light grey.

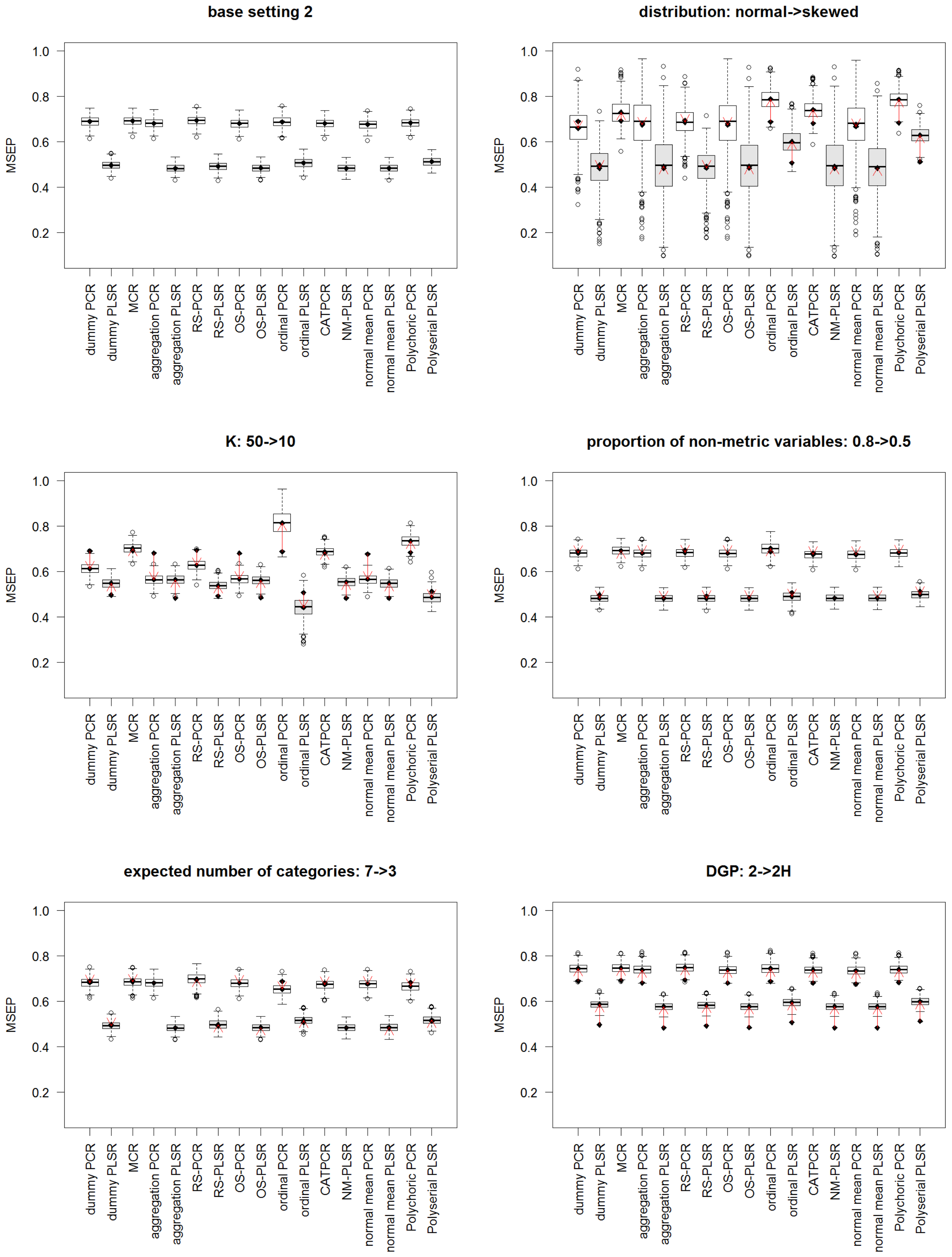
Figure 1 focuses on the comparision of PCA and PLS under two data generating processes. Note that the MSEP-scale of the left and right panel are different. Under DGP 1 both PCA and PLS perform similar, as expected. PLS methods show either little or no advantages compared to PCA. In contrast, under DGP 2 we observe that PLS methods show a clear and significant advantage compared to PCA. Also, under DGP 2 all approaches

Figure 2: MSEP under DGP 1



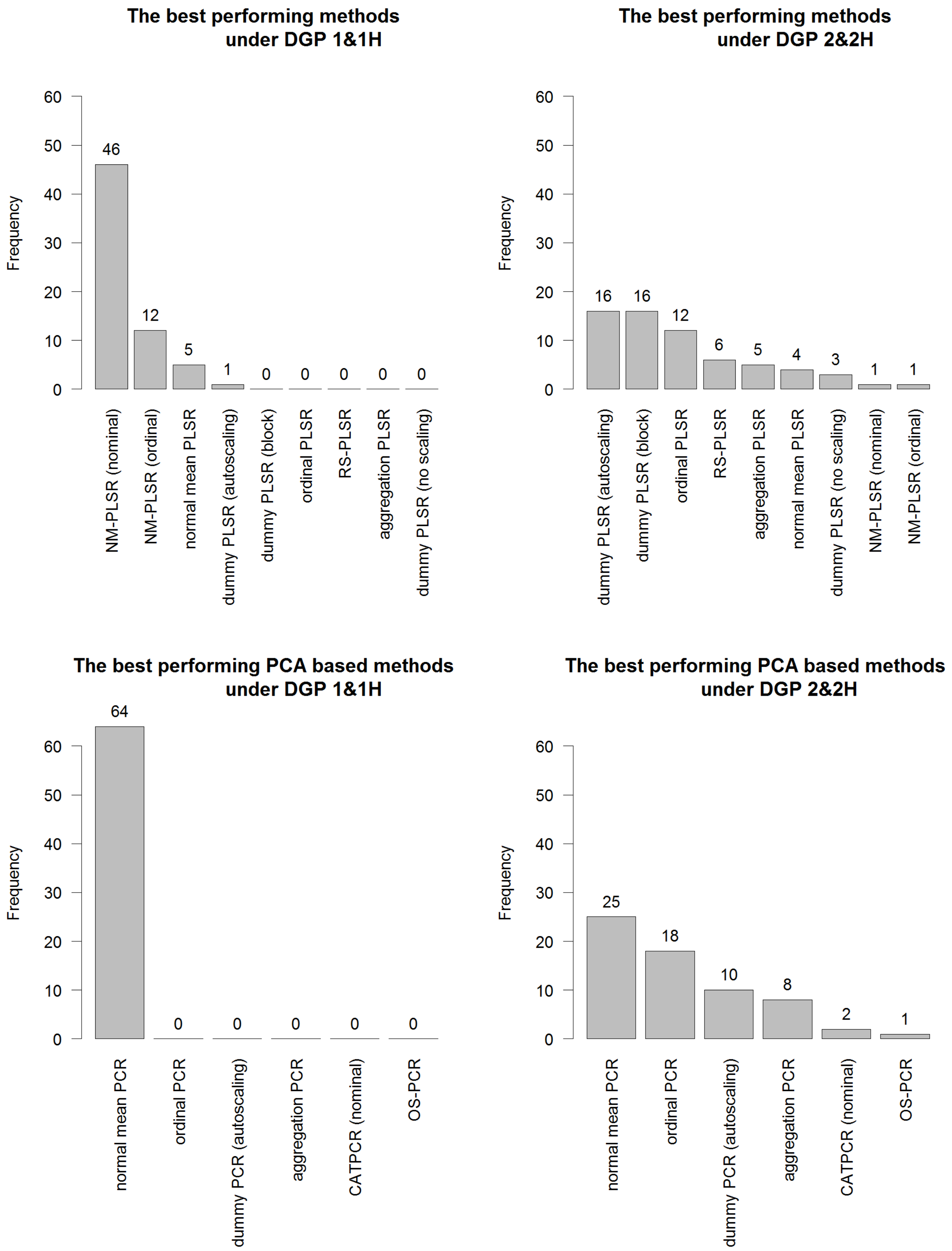
Base setting 1 is used. Red arrows mark changes of the means from the base setting to the respective setting. PCA-based methods are colored white and PLS-based methods light grey.

Figure 3: MSEP under DGP 2



Base setting 2 is used. Red arrows marks changes of the means from the base setting to the respective setting. PCA-based methods are colored white and PLS-based methods light grey.

Figure 4: The absolute frequency of the best perfoming methods over different DGP



for treating non-metric variables perform similar for PCA and PLS, while under DGP 1 several methods show better performance than the others, which we study in much more detail in Figure 2.

Figure 2 shows the performance of various methods under DGP 1. Note that scales on the left middle and right bottom plots are different from the other plots. We focus on *Base setting 1*, shown again in the left top plot and vary one setting at each next plot. The changes of the means from the base setting are marked by red arrows. MCR, RS-PCR, RS-PLSR, ordinal PCR, ordinal PLSR, CATPCA, Polychoric PCR and Polyserial PLSR are not reported, since they performed much worse compared to other methods when the latent variable is skewed and didn't perform good either in other settings as visible in Figure 1. The performance of all remaining methods deteriorates when the true latent variable becomes skewed (right plot in the top row), when the number of the variables decreases (left plot in the middle) and when heterogenous observations are introduced (right plot in the bottom). When the proportion of non-metric variables decreases (right plot in the middle), all methods improve, while the improvement is the most salient for dummy PCR/PLSR. Changes in the expected number of categories (left plot in the bottom) have little impact, except for dummy PCR/PLSR, which noticeably improve with less expected number of categories.

The upper left panel of Figure 4 shows the absolute frequency of best performing (in terms of the average MSEP over Monte Carlo runs) methods out of all 64 settings under DGP 1 and DGP 1H. Even though some methods are not reported in Figure 2 to make the comparison easier, all methods are considered in Figure 4. It is found that NM-PLSR with nominal or ordinal quantification is most often best method followed by normal mean PLSR and dummy PLSR with autoscaling. The lower left panel shows the frequency of best performing PCA-based methods, with normal mean PCR always outperforming other methods. Compared to other methods, dummy coding approach is very attractive in applications due to its simple implementation and interpretation. Therefore, we perform Welch's t-tests to 5% significance level with Bonferroni adjustment (Yandell, 1997, p. 93) to test if NM-PLSR with nominal quantification outperforms dummy PLSR significantly. It turns out, NM-PLSR with nominal quantification is significantly better than dummy PLSR with autoscaling in 59 out of 64 settings. The few settings, where no differences were found, typically have heterogeneity among observations, skewed latent variable and small number of observations. Similarly, normal mean PCR and dummy PCR are tested. It is found that the normal mean PCR significantly outperforms dummy PCR in 62 settings. No differences were found for settings with heterogeneity among observations, skewed latent variable, small sample, many variables and small proportion of non-metric variable.

Figure 3 shows the performance of various methods under DGP 2. When the latent variable is skewed (right top plot), the Monte Carlo variations become large and some methods show deteriorations. With the number of variables decreasing (left plot in the middle row), generally PLS-based methods deteriorate and PCA-based methods improve. But for ordinal PCR/PLSR and polychoric PCR and polyserial PLSR the pattern is opposite. The improvement of ordinal PLSR is so large, that it becomes the best method in this setting. The proportion of non-metric variables (right middle plot) and the expected number of categories (left bottom) do not cause much changes. All methods deteriorate slightly with the heterogeneity among observations (right bottom plot).

The upper right panel of Figure 4 shows the absolute frequency of best performing methods under DGP 2 and DGP 2H. Dummy PLSR with autoscaling and block scaling perform best most frequently followed by ordinal PLSR. The lower right panel shows that normal mean PCR performs most frequently the best among PCA-based methods followed by ordinal PCR and dummy PCR with autoscaling. We performed again Welch's t-tests with Bonferroni corrections as above to test significant differences between methods under all

64 settings. First, ordinal PLSR significantly outperforms dummy PLSR in 16 out of 64 settings. These settings with significant differences typically have high proportion of non-metric variables with few variables. Second, normal mean PCR significantly outperforms dummy PCR with autoscaling in 33 settings. These settings typically have normal distributed latent variable, small number of variables and high proportion of non-metric variables. Third, ordinal PCR significantly outperforms dummy PCR in 20 settings, which typically have normal distributed latent variable.

4 Applications

To demonstrate the performance of PCA and PLS algorithms with non-metric variables on real data, we construct a wealth index, based on the Indonesian Family Life Survey (Strauss et al., 2004) from the year 2000. A wealth index measures household wealth based on the posession of assets and is often used as a proxy for household expenditure. Therefore, we consider the logarithm of the real monthly household expenditure per capita as an outcome variable and aim to find such weights in the wealth index, which provide the best prediction of household expenditure. There are 11 categorical asset variables to build a wealth index. The relationship between wealth and expenditure can differ across observations due to different depreciation rates. Therefore, we consider province, region (kabupaten), destrict (kecamatan) and urban/rural variables to control for heterogeneity. There are 10222 complete observations of households. We use the following empirical model:

$$y_i = T_i\gamma_1 + D_i\gamma_2 + T_i \otimes D_i\gamma_3 + \varepsilon_i,$$

where T_i=$(t_{1i}, ..., t_{Ai})$ contains PCA or PLS scores, D_i is the i-th row of the indicator matrix built from the control variables, $T_i \otimes D_i$ builds the interaction terms between T_i and D_i and γ_1, γ_2 and γ_3 are coefficient vectors of appropriate length.

First, a model selection for the treatment of non-metric variables, the number of scores and control variables is performed. For all treatments of non-metric variables mentioned in Section 2.2, estimated MSEP via 10-fold cross-validation (Mevik and Cederkvist, 2004) is calculated for all possible combinations of the number of scores and control variables. NM-PLSR with 2 scores and province, region and urban/rural variables to control heterogeneity showed the lowest estimated MSEP, closely followed by the PLSR with dummy coding, which we choose due to easier interpretation.

Since dummy coding with autoscaling is used, estimators for $T\gamma_1$ are given by $T\hat{\gamma}_1 = XS^{-\frac{1}{2}}W^*\hat{\gamma}_1 = X\hat{\beta}_1$, where S is a diagonal matrix containing the variance of each column of X and W^* is the PCA or PLS weights in terms of autoscaled regressors. In the following we report $\hat{\gamma}_1$, $\hat{\beta}_1$ and the weights $W = S^{-\frac{1}{2}}W^*$.

Table 1: Coefficient estimates in terms of composite indices and model selection criteria

	$\hat{\gamma}_{1,PCR}$	$\hat{\gamma}_{1,PCR}$	$\hat{\gamma}_{1,PCR}$	$\hat{\gamma}_{1,PCR}$
	$A = 1$	$A = 1, H$	$A = 2$	$A = 2, H$
t_1	0.183***	0.187***	0.183***	0.179***
t_2			−0.055	−0.060
$Adj.R^2$	0.211	0.233	0.222	0.245
$\widehat{MSEP}$	0.446	0.436	0.439	0.429
	$\hat{\gamma}_{1,PLSR}$	$\hat{\gamma}_{1,PLSR}$	$\hat{\gamma}_{1,PLSR}$	$\hat{\gamma}_{1,PLSR}$
	$A = 1$	$A = 1, H$	$A = 2$	$A = 2, H$
t_1	0.211***	0.221***	0.211***	0.210***
t_2			0.103***	0.105***
$Adj.R^2$	0.260	0.281	0.286	0.306
$\widehat{MSEP}$	0.419	0.409	0.404	0.395

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors. The number of scores $A = 1$ or 2. H means that province, region and urban/rural heterogeneity are controlled, which are not reported. $\widehat{MSEP}$ is estimated via 10-fold cross-validation.

Table 1 shows the coefficient estimates $\hat{\gamma}_1$ for PCA or PLS and model selection statistics. Apparently, the PCR-based model has quite low adjusted R^2 and the model hardly improves by adding an additional score. In all models the estimated coefficients by t_1 are significant, but the coefficients by t_2 are not significant. In contrast, the PLSR-based model improves significantly, if one more score is added. Thereby, all coefficients are highly significant. Taking heterogeneity into account brings similar gains to the PCR and PLSR. The PLSR with two scores and heterogeneity control shows that with increasing wealth, measured by the first and second score, expenditure is predicted to increase. The PCR shows analogous results, except that the second score is not significant.

Table 2 shows the coefficient estimates in terms of the variables building the scores and weights. The coefficient estimates of the PCR and PLSR under our favored setting, i.e., with heterogeneity control and two scores, show strong differences, while the PLSR coefficients are better in terms of prediction as shown in Table 1. The PCR and PLSR coefficients of owning farm land and cooking with kerosene have opposite signs. The PLSR emphasizes refrigerators, owning house and buildings, using mineral water as drinking water, using public toilet and all variables related to cooking, while electricity, piped, surface, rain, basin water, toilet without septank and communal toilet are less important compared to the PCR. Analogously, PLS and PCA weights show strong differences, which show the weights better suited for the prediction of household expenditure. The first PLS weights emphasize owning non-farm land, using mineral water as drinking water, public toilet, cooking with electricity and don't cook, while owning farm land, using communal toilet and cooking with kerosine are less important compared to the first PCA weight. The second PLS and PCA weights show more drastic differences, where more than half of the variables having weights of opposite signs. Introducing the second score brings larger changes in coefficient estimates in the PLSR compared to the PCR, which is not surprizing given that the PCR coefficient estimate in terms of the second score in Table 1 is not significant. We see large differences between the PLSR with one and two scores in electricity, owning farm land, using surface, basin and mineral water as drinking water, toilet without septank, public toilet, cooking with electricity and kerosene and don't cook. The PCR with one and two score shows moderate differences in owning house and non-farm land, communal and public toilet, cooking with kerosene and don't cook.

Table 2: PCR and PLSR coefficients in terms of the variables building the composite indices and weights

	$\hat{\beta}_{PCR}$	$\hat{\beta}_{PLSR}$	$\hat{\beta}_{PCR}$	$\hat{\beta}_{PLSR}$	$w_{1,PCA}$	$w_{1,PLS}$	$w_{2,PCA}$	$w_{2,PLS}$
	$A=1,H$	$A=1,H$	$A=2,H$	$A=2,H$				
electricity	0.171***	0.139***	0.150***	0.044***	0.915	0.629	0.227	−0.835
television	0.114***	0.126***	0.135***	0.108***	0.612	0.568	−0.428	−0.111
refrigerators	0.152***	0.239***	0.195***	0.312***	0.812	1.081	−0.824	0.809
vehicle	0.060***	0.057***	0.083***	0.028***	0.323	0.256	−0.419	−0.243
own: house	−0.067***	−0.094***	−0.016**	−0.125***	−0.357	−0.425	−0.797	−0.337
own: buildings	0.080***	0.122***	0.096***	0.146***	0.426	0.551	−0.329	0.286
own: non-farm land	0.004	0.030***	0.041***	0.058***	0.023	0.137	−0.606	0.278
own: farm land	−0.089***	−0.047***	−0.047***	0.041***	−0.479	−0.215	−0.640	0.815
water: piped	0.107***	0.095***	0.105***	0.034***	0.571	0.431	−0.041	−0.532
water: well	−0.048***	−0.070***	−0.053***	−0.070***	−0.257	−0.314	0.124	−0.035
water: surface	−0.132***	−0.101***	−0.108***	−0.023	−0.708	−0.455	−0.311	0.691
water: rain	−0.046***	−0.031	−0.040**	0.002	−0.248	−0.139	−0.067	0.296
water: basin	−0.092***	−0.071***	−0.089***	−0.014	−0.493	−0.321	0.019	0.505
water: mineral	0.102***	0.261***	0.092***	0.428***	0.547	1.177	0.095	1.716
toilet: septank	0.139***	0.158***	0.150***	0.136***	0.743	0.713	−0.289	−0.130
toilet: no septank	−0.070***	−0.057***	−0.056***	−0.014	−0.374	−0.257	−0.186	0.379
toilet: communal	−0.019***	−0.004	−0.074***	0.028	−0.103	−0.019	0.928	0.302
toilet: public	−0.009*	−0.057***	−0.050***	−0.119***	−0.050	−0.257	0.675	−0.619
toilet: field	−0.127***	−0.157***	−0.126***	−0.157***	−0.677	−0.708	0.077	−0.082
cooking: electricity	0.035**	0.210***	0.039**	0.458***	0.190	0.948	−0.076	2.460
cooking: gas	0.137***	0.239***	0.201***	0.317***	0.732	1.079	−1.161	0.863
cooking: kerosene	0.077***	0.020***	0.021*	−0.051***	0.413	0.092	0.880	−0.671
cooking: wood, coal	−0.156***	−0.171***	−0.117***	−0.171***	−0.838	−0.772	−0.548	−0.082
cooking: don't cook	0.042***	0.259***	−0.031***	0.575***	0.223	1.172	1.171	3.126

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors. The number of scores $A = 1$ or 2. H means that province, region and urban/rural heterogeneity are controlled, but not reported. As base categories "water: other", "toilet: other" and "cooking: other" are excluded.

5 Conclusions

We have reviewed various treatments of non-metric variables in PCA and PLS algorithms. The results of the simulation study suggest the following. First, PLS-based methods are to prefer in practice, since, independent of true data generating process, PLS performs either as good as PCA or significantly outperforms it. Second, under considered data generating processes, NM-PLSR performs best under DGP 1&1H, while dummy PLSR is to prefer under DGP 2&2H. Ordinal PLSR shows good performance in a few occasions under DGP 2&2H. Third, normal mean PCR showed most often the best performance, followed by ordinal and dummy PCR. Finally, ignoring heterogeneity among observations leads to a deterioration for all methods and settings.

As an application wealth indices to predict household expenditure have been considered. The number of scores and variables to control heterogeinity are selected simultaneously, which bring gains in prediction performance and large changes to coefficients. The weights and coefficients of PLSR and PCR differ drastically, while the weights and coefficients of PLSR turn out to be better for the prediction.

References

Barro, R. J. (1989). Economic growth in a cross section of countries. *National Bureau of Economic Research.* w3120.

Branisa, B., Klasen, S., and Ziegler, M. (2013). Gender inequality in social institutions and gendered development outcomes. *World Development*, 45:252–268.

Central Bureau of Statistics (CBS) Kenya, Ministry of Health (MOH) Kenya, and ORC Macro (2004). Kenya Demographic and Health Survey 2003. url = http://www.measuredhs.com/. CBS, MOH, and ORC Macro, Calverton, Maryland.

Chin, W. W., Marcolin, B. L., and Newsted, P. R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a monte carlo simulation study and an electronic-mail emotion/adoption study. *Information systems research*, 14(2):189–217.

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory System*, 18:251–263.

Dreher, A. (2006). Does globalization affect growth? Evidence from a new index of globalization. *Applied Economics*, 38(10):1091–1110.

Filmer, D. and Pritchett, L. H. (2001). Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of India. *Demography*, 38(1):115–132.

Greenacre, M. (2010). *Correspondence Analysis in Practice.* Chapman and Hall/CRC.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417–441.

IBM SPSS Statistics (2013). Categorical Principal Components Analysis (CATPCA). url=http://www-01.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_catpca_obj-func-opt_opt.htm.

Keun, H. C., Ebbels, T., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2003). Improved analysis of multivariate data by variable stability scaling: application to nmr-based metabolic profiling. *Analytica chimica acta*, 490(1):265–276.

Kolenikov, S. and Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?. *Review of Income and Wealth*, 55(1):128–165.

Meulman, J. (2000). Optimal scaling methods for multivariate categorical data analysis. *Leiden: Leiden University*, 12.

Mevik, B.-H. and Cederkvist, H. R. (2004). Mean squared error of prediction (msep) estimates for principal component regression (pcr) and partial least squares regression (plsr). *Journal of Chemometrics*, 18(9):422–429.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

Naes, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data. *Communications in Statistics-Simulation and Computation*, 14(3):545–576.

Nardo, M., Saisana, M., Saltelli, A., and Tarantola, S. (2005). Tools for composite indicators building. European Comission, Ispra.

Niitsuma, H. and Okada, T. (2005). Covariance and pca for categorical variables. In *Advances in Knowledge Discovery and Data Mining.*, pages 523–528. Springer, Berlin Heidelberg.

Russolillo, G. (2009). *Partial Least Squares Methods for Non-Metric Data.* PhD thesis, Università degli Studi di Napoli Federico II.

Rutstein, S. O. and Johnson, K. (2004). The DHS wealth index. ORC Macro, MEASURE DHS.

Sachs, J. D. and Warner, A. M. (1997). Sources of slow growth in african economies. *Journal of African economies*, 6(3):335–376.

Saisana, M. and Tarantola, S. (2002). State-of-the-art report on current methodologies and practices for composite indicator development. EUR 20408 EN, European Commission-JRC: Italy.

Strauss, J., Beegle, K., Sikoki, B., Dwiyanto, A., Herawati, Y., and Witoelar, F. (2004). The third wave of the Indonesia Family Life Survey (IFLS3). url = http://www.rand.org/labor/FLS/IFLS.html. Overview and field report. NIA/NICHD.

Tenenhaus, M. and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50(1):91–119.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In *Research* *papers in statistics*. Wiley, New York.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Yandell, B. S. (1997). *Practical data analysis for designed experiments.*, volume 39. CRC Press.