Partial Least Squares and Principal Component Analysis with Non-metric Variables for Composite Indices

Dissertation

zur Erlangung des wirtschaftswissenschaftlichen Doktorgrades der

Wirtschaftswissenschaftlichen Fakultät der Universität Göttingen

vorgelegt von

Jisu Yoon

aus

Cheonan, Südkorea

Göttingen, 2015

Erstgutachter: Prof. Tatyana Krivobokova, Ph.D.

Zweitutachter: Prof. Stephan Klasen, Ph.D.

Tag der Disputation: 24. April 2015

Contents

Acknowledgements

I thank my advisors, Tatyana Krivobokova and Stephan Klasen, for their wonderful supervision and advices. I thank Axel Dreher for his insights and helps, and for joining the committee.

University of Göttingen has provided me good opportunities to build a background in economics and to prepare for a career. I am grateful to the German Research Foundation (DFG) for the funding during my Ph.D.. The scholarship from the National Institute for International EDucation (NIIED) helped me to begin my study and research in Germany.

During my Ph.D., I shared good times with my friends and colleagues from my office and the research training group, Globalization and Development (GLaD), and with many others. Finally, I thank everyone who helped me on my way to here.

List of Abbreviations

BMI Body Mass Index

CATPCA Categorical Principal Component Analysis

CIRI Cingranelli-Richards Human Rights Dataset

CPI Corruption Perception Index

DGP Data Generating Process

DHS Demographic Health Survey

EAC East Asia and Pacific

ECA Europe and Central Asia

FA Factor Analysis

FDI Foreign Direct Investment

GDP Gross Domestic Product

LAC Latin America and Caribean

MCA Multiple Correspondence Analysis

MENA Middle East and North Africa

NIPALS Non-linear Iterative PArtial Least Squares

NM-PLSR Non-Metric Partial Least Squares Regression

OECD Organisation for Economic Co-operation and Development

PCA Principal Component Analysis

PCR Principal Component Regression

PLS Partial Least Squares

PLSR Partial Least Square Regression

PPP Purchasing Power Parity

SA South Asia

SIGI Social Institutions and Gender Index

SSA Sub-Saharan Africa

List of Figures

List of Tables

Introduction

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). Composite indices are often used to measure latent phenomena or to summarize complex information in a small number of variables. For example, the Corruption Perception Index (CPI; Transparency International, 2013) quantifies the level of corruption in various countries. Survey variables on various types of people with different foci of questions and various expert opinions are aggregated to build this index. The CPI can be used to generate a cross country ranking (Transparency International, 2013), or to research the relationship between curruption and foreign direct investment (FDI; Habib and Zurawicki, 2002) or gender inequality (Branisa et al., 2013). The KOF Index of Globalization (Dreher, 2006) quantifies globalization across countries, which is composed of economic, social and political globalization. Each facet of globalization is measured as a linear combination of relevant correlates, e.g., trade in percent of GDP, number of McDonald's restaurants per capita and participation in the U.N. security council missions. This index is used to generate a cross country ranking (KOF Swiss Economic Institute, 2013), or to study the relationship between globalization and growth (Dreher, 2006; Rao et al., 2011) or human rights (Potrafke, 2014).

It is crucial to choose correct weights for the variables that build a composite index. There are several approaches to assign weights available in the literature. Apart from subjective and non-data driven ways to assign weights, Principal Component Analysis

(PCA; Filmer and Pritchett, 2001) is a popular approach, which determines weights, so that the largest variations in variables are emphasized in the resulting composite index. Factor Analysis (FA; Sahn and Stifel, 2000) and Multiple Correspondence Analysis (MCA; Booysen et al., 2008) determine weights similarly. PCA may perform poorly if the largest variations in variables are not informative, which occurs when observed variables contain large measurement errors or variations coming from other latent variables. For example, one may try to measure cross country corruption using survey variables. The value of the survey variables may not only be influenced by corruption, but also the quality of journalism, which report the corruption in the country to the public, or the attitude of surveyees. To quantify the level of globalization, one may use the number of McDonald's restaurants in a country. But this variable is also influenced by the presence of competitors such as Wendy's Burger or Burger King. In some countries burger bread may not be popular because of low quality wheat caused by climate and land conditions. If the largest variations in variables come from such measurement errors or irrelevant latent factors, PCA will measure something different than the concept that a composite index is supposed to capture.

The main contribution of this work is applying Partial Least Squares (PLS; Wold, 1966b) to assign weights in composite indices to avoid the aforementioned problem of PCA. PLS assigns weights, so that variables showing high covariance with respect to particular outcome variables are emphasized in the composite index. Consequently, PLS weights draw information from the structural relationship between outcome variables and a latent concept, which is manifested to observed variables. For example, if one expects that globalization influences economic growth significantly, one can build a globalization index with weights, which emphasize variables covarying with economic growth. If globalization actually has significant influence on growth and the observed variables contain certain amount of variations from globalization, PLS will measure globalization better than PCA, especially when the largest variations in variables are not related to globalization. Using

PLS has the following additional advantages. First, a composite index using PLS often leads to a better prediction for a certain outcome variable than a composite index using PCA. As a result, one can generate a composite index particularly relevant for the outcome variable. This procedure is especially useful when the latent concept of interest is multidimensional. For example, globalization may have several dimensions, each of them relevant for economic growth, human rights and inequality. PLS can generate composite indices, each tailored to one of these outcome variables. Second, a comparison between PLS and PCA weights shows which variables are relevant for the prediction of a particular outcome variables. On the other hand, PLS had a caveat that coefficients in a regression analysis cannot be interpreted as causal relationship, because a composite index based on PLS already contains information from the outcome variable.

In practice variables that enter a composite index are often non-metric (ordinal and nominal). For example, the level of violence against women and the discrimination against women in terms of access to loans are measured in ordinal scale, which are used to build a composite index regarding gender inequality (Branisa et al., 2013). PCA and PLS can be applied on non-metric variables only with a special treatment. As the second contribution of this work, we review various PCA and PLS algorithms for non-metric variables available in the literature, which have different motivations and assumptions on data generating processes (DGPs). This study provides extensive simulation studies to compare the performance of the methods under typical DGPs and make recommendations for practitioners. In real data applications, we select appropriate methods for non-metric variables based on model selection criteria. The methods under consideration are dummy coding (Filmer and Pritchett, 2001), multiple correspondence analysis (MCA; Greenacre, 2010), the aggregation method (Saisana and Tarantola, 2002), the regular simplex method (Niitsuma and Okada, 2005), 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). Additionally, we consider three methods from Kolenikov and Angeles (2009), the normal mean coding, ordinal PCA/PLS and polychoric PCA, and modify polychoric PCA in a PLS context, which we call polyserial PLSR.

This dissertation is composed of three essays, which are summarized in the followings.

Essay 1: Composite Indices Based on Partial Least Squares

This essay generates three composite indices, which are two wealth indices and a globalization index, and compares and selects the treatment of non-metric variables in PCA and PLS based on a simulation study and model selection criteria.

First, we compare composite indices based on PCA and PLS with various treatments of non-metric variables in terms of prediction performance using simulation studies, when we use composite indices to summarize variables. The results show that composite indices based on PLS outperform composite indices based on PCA and dummy coding performs satisfactorily compared to more sophisticated statistical procedures. We favor dummy coding not only because it performs good, but also it is easy to implement and interpret.

We consider three applications. First, the Body Mass Index (BMI) of adult population in Kenya is predicted by a wealth index. A wealth index measures household wealth typically as a linear combination of household asset possessions. The BMI is expected to be influenced by wealth (Wittenberg, 2013), while low wealth may lead to undernutrition or overweight. Second, household expenditure in Indonesia is predicted by another wealth index. A wealth index is often used to proxy household expenditures and the appropriate weights for this task is an important question. Third, economic growth is predicted by the KOF Index of Globalization (Dreher, 2006) with new weighting schemes. Globalization influences economic growth (Dreher, 2006; Rao et al., 2011) and we try to find the weights better predicting economic growth.

Coherent with the simulation study, the results indicate that composite indices using PLS show better prediction performance and fitting than composite indices using PCA. Model selection statistics support the use of dummy coding as the treatment of non-metric variables. PLS and PCA generate substantially different weights and coefficients, which can be compared to find out the relevant variables in a composite index for the prediction of a particular outcome variable. More wealth predicts higher BMI and more household expenditure, while globalization predicts higher economic growth.

Essay 2: Treatments of Non-metric Variables in Partial Least Squares and Principal Component Analysis

In this essay, the treatments of non-metric variables in PCA and PLS are reviewed in more detail followed by extensive simulation studies to make recommendations under typical DGPs and a wealth index application.

After reviewing the treatments of non-metric variables in PCA and PLS in detail, simulation studies follow. The simulation design is changed, so that a composite index is used to capture a latent variable. We compare the performance of PCA- and PLS-based composite indices with the treatments under various DGPs, which are selected considering typical DGPs in practice. Based on the simulation results, we provide recommendations for the treatments under various DGPs. Composite indices based on PLS are either superior or as good as composite indices based on PCA. PLS with dummy coding is often attractive when the variables building the composite index contain little variations from the latent variable of interest. Other methods, such as NM-PLSR, PCA with normal mean coding, ordinal PCA and PLS, show good performance in certain conditions.

As our application we revisit the wealth index to predict household expenditure in Indonesia. We perform a model selection in terms of the number of scores, control variables and the treatments of non-metric variables at the same time to improve the prediction performance. Model selection statistics suggest again that PLS outperforms PCA and dummy coding is an attractive treatment for non-metric variables. Using two scores and introducing control variables bring noteworthy gains, with which PLS and PCA show even larger differences in terms of weights and coefficients. Wealth again predicts higher expenditure.

Essay 3: An Application of Partial Least Squares to the Construction of the Social Institutions and Gender Index (SIGI) and the Corruption Perception Index (CPI)

This work focuses on measuring gender inequality and corruption using composite indices based on PLS and PCA and studies the effects of gender inequality on female education, child mortality, fertility and corruption.

Gender inequality is believed to have negative effects on the development of the society (Sen, 1999) in addition to the deprivation of women from basic rights. Branisa et al. (2013) have created the Social Institutions and Gender Index (SIGI) to measure social institutional aspects of gender inequality. The SIGI is used to explain several gender outcomes, i.e., female education, fertility, child mortality and corruption, measured by the Corruption Perception Index (CPI; Transparency International, 2013). Branisa et al.'s weighting scheme involves arbitrary judgements and could be improved to predict the outcome variables better. Therefore, we change the weighting scheme of the SIGI to PCA and PLS and redo the empirical exercises to explain the gender outcomes.

The results show that PLS and PCA again generate substantially different weights and coefficients. For female education and child mortality, Partial Least Square Regression (PLSR) shows better prediction performance than Principal Component Regression

(PCR) and we could find out the variables which particularly matter for the prediction of the respective outcome variables. Both PCR and PLSR find that high gender inequality leads to more fertility and high corruption, while for female education and child mortality the PLSR find significant relationship, while the PCR doesn't. The significant coefficient estimates from the PLSR cannot be interpreted as a causal relationship because PLS weights contain information from the outcome variable. But since PLSR is more robust against measurement errors, we can suspect that the insignificant coefficient estimates from the PCRs are caused by measurement errors. Dummy coding is selected as the treatment of non-metric variables based on estimated prediction performance, because it generally works well, albeit not always the best, and is easy to implement and interpret. Additionally, we take a close look on the CPI. The CPI is a composite index using a simple average as the weighting scheme. If all variables building the CPI are not equally important, a simple average is not the best way of aggregating. Therefore, we use PCA and PLS to generate weights for the CPI. We select the variables differently to drop low quality data and not to emphasize certain variables without good reasons. With these new CPIs we find again that gender inequality leads to more corruption.

Concluding Remarks

In this study, we use both PCA and PLS to generate composite indices, while giving a special attention on the treatments of non-metric variables. We review the treatments available in the literature and compare them by means of simulation studies. The simulation studies suggest that composite indices based on PLS outperform composite indices based on PCA. PLS with dummy coding is often attractive in terms of performance, the ease of implementation and interpretation. Additionally, we check the performance of the treatments in real data analyses using cross-validations. For the majority of our applications, PLS with dummy coding shows good performance and PLS-based composite indices outperformed PCA-based composite indices. Consequently, we could generate composite indices tailored to particular outcome variables, and the comparison between PLS and PCA weights and coefficients showed which variables in a composite index were particularly relevant for a certain outcome variable.

Chapter 1

Composite Indices Based on Partial Least Squares

with Stephan Klasen, Axel Dreher and Tatyana Krivobokova

Abstract

In this paper, we compare Principal Component Analysis (PCA) and Partial Least Squares (PLS) methods to generate weights for composite indices. In this context we also consider various treatments of non-metric variables when constructing such composite indices. Using simulation studies we find that dummy coding for non-metric variables yields satisfactory performance compared to more sophisticated statistical procedures. In our applications we illustrate how PLS can generate weights that differ substantially from those obtained with PCA, increasing the composite indices' predictive performance for the outcome variable considered.

1.1 Introduction

Composite indices are often used in economics to summarize complex information into a single number with the aim to simplify more complex phenomena or for comparative and ranking purposes. 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). That is, a composite index is a special linear combination of several variables, related to a certain concept. An example of a composite index aiming to capture a latent variable is the wealth index commonly used to proxy for income in Demographic and Health Surveys (Rutstein and Johnson, 2004), while composite indices used for aggregation and ranking purposes include the Summary Innovation Index (DG Enterprise, 2001). In regression models such indices lessen the multicollinearity problem and can be easier to interpret than original variables.

Naturally, the quality of a composite index depends on the choice of weights, for which the literature provides several possibilities. Apart from the researcher's subjective choice, weights based on the variance-covariance structure of variables are most widely used. Principal Component Analysis (PCA; e.g. Filmer and Pritchett, 2001), Factor Analysis (FA; e.g. Sahn and Stifel, 2000) and Multiple Correspondence Analysis (MCA; e.g. Booysen et al., 2008) are popular methods to set weights in a composite index. All of these techniques are meant to extract the largest variation in the variables building a composite index. However, often the largest variation is not related to a response variable, which one wishes to explain using the composite index. Therefore, we propose to apply Partial Least Squares (PLS; Wold, 1966b) to build composite indices in order to find the weights for the variables that are most relevant for a particular response variable. To put it simply, while PCA and related methods find the weights which maximize the covariance of the vector of independent variables, PLS weights maximize the covariance between covariates and a certain response variable. In consequence, PLS extracts factors relevant to a particular response variable, instead of build an 'all-purpose' index. Therefore, we see several advantages in the application of PLS when constructing composite indices. First, using PLS weights designed for a certain outcome variable should improve the prediction of this variable via the resulting composite index. Such composite indices can be used for prediction and as diagnostic tools that shows which indicators included in a composite index are particularly important for the outcome variable, thus adjusting the composite index to the particular problem at hand. Second, comparing PCA- and PLS-based weights, one can infer which variables in the composite index are particularly important for a certain response. Third, by definition one can expect PLS to be more robust than PCA in the presence of measurement errors. On the other hand, a composite index based on PLS has a caveat, that one cannot infer a causal relationship from regression analysis, since the composite index already contains information from the response variable. It seems to be possible to circumvent this problem using a simple two step procedure to make PLS scores exogenous in a regression analysis, which is not the focus of this study.[1]

Similar to PLS, weighting schemes based on regression (Ravallion, 2012a,b) consider the relationship between a particular response variable and covariates. But weighting based on regression is vulnerable against multicollinearity of covariates, because it can involve an inversion of a (nearly) singular matrix, whereas PLS doesn't. PLS is a technique for multicollinear data (Naes and Martens, 1985).

Many variables used to build composite indices, especially in economic applications, are non-metric, which hinders direct application of PLS and PCA methods, because PLS and PCA are primarily developed for continuous variables. Therefore, in this work we also discuss and compare in simulations the prediction performance of various treatments of non-metric variables in PCA and PLS available in the literature. It turns out that using

[1] Consider that we have two outcome variables, Y_{train} and Y_{test}, whereby $Y_{test} = X\beta_{test} + \varepsilon_{test}$ and $\mathbb{E}(\varepsilon_{test}|Y_{train}, X) = 0$. We build a PLS score using a relationship between Y_{train} and X, so that $S_1 = XX^tY_{train}/\left\|X^tY_{train}\right\|$. Obviously, it follows that $\mathbb{E}(\varepsilon_{test}|S_1) = 0$, so that a causal interpretation of Y_{test} on S_1 is possible.

dummy coding typically provides very good predictions and is easy to interpret.

To illustrate the performance of PCA- and PLS-based composite indices we consider wealth and globalization indices. A wealth index aims to describe household wealth based on the possession of certain asset variables. This index is particularly attractive in the context of developing countries, since conventional measurements such as income or consumption expenditures are hard to obtain or of low quality (for other advantages of wealth indices see Rutstein and Johnson, 2004). Therefore, in this work we build wealth indices based on the Kenyan Demographic Health Survey of 2003 (Central Bureau of Statistics (CBS) Kenya et al., 2004) and on the Indonesian Family Life Survey from the year 2000 (Strauss et al., 2004). In the Kenyan example we choose the respondent's BMI to be the response variable that we seek to correlate with the wealth index. In the case of Indonesia, we choose household expenditures as the response variable to assess which weights of the wealth index provide a particularly good proxy for expenditures. The globalization index we chose for our analysis is the KOF Index of Globalization (Dreher, 2006), which we relate to economic growth. The index aims to quantify the phenomenon of globalization, which is defined as the process of creating connections between actors at multicontinental distances, which are mediated through a variety of flows including people, information and ideas, capital and goods (based on Clark, 2000; Norris, 2000; Keohane and Nye, 2000). The data for this index come from Dreher (2006) and economic growth is used as an outcome variable to create a version of the Globalization Index whose weights are particularly closely related to growth.

The paper is organized as follows. In Section 1.2 we review basic principles of PLS and PCA, various treatments of non-metric variables for these algorithms and conduct a simulation study. Section 1.3 presents the analysis of the three data sets and the indices we obtain, while we conclude in Section 1.4.

1.2 PCA and PLS with Non-metric Variables

1.2.1 PCA and PLS algorithms

Let X be a $n \times k$, $k < n$, centered matrix, which contains n observations of k-dimensional vector of (metric) covariates. PCA is a natural way to reduce the covariate dimension k and avoid collinearity problems in a linear regression model

$$Y = X\beta + \varepsilon, \tag{1.1}$$

for $Y = (y_1, \ldots, y_n)^t$, $\beta = (\beta_1, \ldots, \beta_k)^t$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$, with $E(\varepsilon) = 0_n$, $\text{cov}(\varepsilon) = \sigma^2 I_n$. The first principal component equals to such a linear combination of covariates, that has the maximum empirical covariance, that is $P_1 = Xu_1$, where

$$u_1 = \arg\max_{\|u\|=1} u^t X^t X u$$

is the k-dimensional first eigenvector of $X^t X$, which corresponds to the maximum eigenvalue. Further principle components are found from the same maximization problem under the orthogonality constraint, that is

$$u_i = \arg\max_{\|u\|=1} u^t X^t X u, \text{ subject to } u_i \perp \ldots \perp u_1, \;\; i = 2, \ldots, k,$$

which corresponds to the ith eigenvector of $X^t X$.

The PLS algorithm follows a similar paradigm, except that the squared empirical covariance between X and Y is maximized, that is $S_1 = X\omega_1$ with

$$\omega_1 = \arg\max_{\|\omega\|=1} \omega^t X^t Y Y^t X \omega \propto X^t Y$$

and further ω_i solving the same optimization problem, again subject to mutual orthogonality of all $\omega_i, \ldots, \omega_1$.

Composite indices are typically built using only the first component, we therefore define a PCA-based composite index as $P = Xu_1$ and a PLS-based composite index by $S = X\omega_1$. This makes the difference between both indices apparent: PCA-based indices use the first eigenvector of X^tX as weights, while PLS-based indices have weights X^tY.

Finally we note, that PCA and PLS depend on the scaling of variables (Wold et al., 2001; Keun et al., 2003). Autoscaling is commonly used which not only centers each variable, but also scales it to unit variance.

1.2.2 Non-metric Variables in PCA and PLS

Composite indices often include non-metric variables. In the following we discuss several approaches available in the literature to perform PCA and PLS in the presence of non-metric variables. The outcome variable is always metric.

The first approach is to transform each category of a non-metric variable to a variable and PCA or PLS is performed as usual. This approach is used in **dummy coding** (Filmer and Pritchett, 2001), **multiple correspondence analysis** (MCA; Greenacre, 2010), the **aggregation method** (Saisana and Tarantola, 2002) and the **regular simplex method** (Niitsuma and Okada, 2005). **Dummy coding** just translates each category of a non-metric variable into a dummy variable. Consequently, each non-metric variable is transformed to an indicator matrix, where one category may be omitted for the ease of interpretation. **MCA** extends simple dummy coding in that the columns of the obtained indicator matrix are weighted so that categories with many incidences and categories with few incidences are equally important. An **aggregation method** can be used for observations belonging to clusters, replacing each dummy variable in the indicator matrix with the cluster level average. The **regular simplex method** transforms each unique category of a non-metric variable to the corresponding vertex coordinate of a regular simplex. The dimension of the regular simplex is selected so that the number of vertices and the number of unique categories are equal.

Another approach is to scale each unique category of non-metric variables. Afterwards, scaled variables are considered to be metric and PCA or PLS are applied as usual. This technique is used in 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). These methods involve an optimization with respect to category values. The **optimal scaling method** maximizes the sum of variances of the scaled variables. **NM-PLSR** maximizes the covariance between the first PLS score and the outcome variable. **CATPCA** maximizes the sum of variances of the PCA scores. The optimizations in all three methods require appropriate constraints for a solution to exist.

We also mention **polychoric PCA** (Kolenikov and Angeles, 2009), which assumes that each observed ordinal variable is generated by a normally distributed latent process, which is discretized at unobserved thresholds. **Polychoric PCA** is performed on the variance-covariance matrix of latent variables, obtained according to the assumed data generating process. Autoscaling is applied to the variables building the scores. **Normal mean coding** is a related method based on the same distributional assumption as polychoric PCA from the same authors, which scales each category value of an ordinal variable as the group mean of the latent process. There is an approach to use polychoric and polyserial correlation in the context of PLS (Cantaluppi, 2012), but this paper restricts its attention to a simple method in analogy to polychoric PCA, which is named as **polyserial PLSR**. We apply autoscaling to regressand and regressors and calculate the polyserial or Pearson correlation between them. The correlation vector is standardized to unit length, which is used as the weight vector to extract the PLS score.

Ordinal PLS or **PCA** treats ordinal variables as numerical variables and apply PLS or PCA respectively. These methods are not recommended since the scaling of an ordinal variable usually contains large errors, but it can serve as a reference for other methods.

In the following we compare various treatments of non-metric variables in PCA and PLS in a simulation study in terms of prediction performance. In the i-th run out of $M = 500$ Monte Carlo runs, data are generated according to model (1.1)

$$Y_i = X_i\beta + \varepsilon_i,\ i = 1, \dots, M,$$

where the number of observations is $n = 5000$ and the covariate dimension is $k = 50$. Regressors are simulated from the standard multivariate normal distribution. The correlation between each pair of variables is generated from the uniform distribution on $[-0.999, 0.999]$. Each regressor is divided by its standard deviation, so that the variance equals 1. We generate β once from the standard normal distribution, which does not change over Monte Carlo simulations. The error term is generated from $\varepsilon_i \sim \mathcal{N}(0_n, 9I_n)$. If a variable is set to be a non-metric variable, it is discretized. To have m_j number of unique categories for the j-th variable, $m_j - 1$ thresholds are generated from the uniform distribution on $[0, 1]$. Next, the empirical CDF of the variable is calculated and we divide the quantiles to m_j number of segments using the thresholds. The variable values corresponding to the lowest segment to the highest segment receive integer values from zero to $m_j - 1$ respectively. The number of unique categories m_j is generated once and does not change over Monte Carlo runs. Thereby, m_j is generated from the Poisson distribution with mean λ and 2 is added to guarantee that each variable has at least two unique values. For example, if the expected number of unique categories is set to be 2.5, $m_j = m_j^* + 2$ where $m_j^* \sim Poi(\lambda = 0.5)$. Most of the treatments imply particular scalings for non-metric variables, which we do not change. But for dummy coding three types of data scalings are considered: no scaling, auto scaling and block scaling. For block scaling, the sum of variances from the dummy variables from each non-metric variable is set to be one.

We consider four scenarios:

		Expected number of unique categories	
		2.5	10.5
Non-metric	10%	Scenario 1	Scenario 3
variables	50%	Scenario 2	Scenario 4

That is, under Scenario 1 matrix X contains 10% of non-metric variables and the number of unique categories over all categorical variables is 2.5 in the mean and so on.

Prediction performance is measured by the average of the mean squared error of prediction (MSEP) defined by

$$MSEP = \frac{1}{Mn}\sum_{i=1}^{M}(X_i\beta_i - U_i\hat{\gamma}_i)^t(X_i\beta_i - U_i\hat{\gamma}_i)$$

The columns of U include the intercept and the first score, that is, $U = (\underline{1}_n, P)$ for PCA and $U = (\underline{1}_n, S)$ for PLS, where $\underline{1}_n = (1, \dots, 1)^t$ is a n-dimensional vector of ones and P and S as defined in Section 1.2.2. The coefficient vector $\hat{\gamma}_i$ is the OLS coefficient estimates of Y_i on U_i.

Table 1.1: Prediction performance in terms of MSEP

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
dummy PCR (autoscaling)	71.09	71.73	70.93	71.24
dummy PLSR (autoscaling)	10.72	11.66	11.49	14.99
polychoric PCR	70.91	71.09	73.25	73.89
polyserial PLSR	11.59	13.64	16.73	21.66
CATPCR	70.93	71.16	70.87	71.07
NM-PLSR	15.50	35.27	14.81	33.36

Table 1.1 reports the simulations results. First, we observe that PLS-based methods perform better than PCA-based ones in all settings. Furthermore, PCA-based methods do not differ much from one to the other in terms of performance. Under PLS-based methods dummy coding with autoscaling performs best followed by polyserial PLSR and NM-PLSR. Second, the performance deteriorates with increases in the proportion of non-metric variables, while NM-PLSR shows the largest deterioration. Third, increasing the expected number of categories usually has little influence, except for polyserial PLSR and dummy PLSR we see notable deterioration. For all scenarios we also ran simulations for other methods discussed in Subsection 1.2.2 and found the following results. Principal Component Regressions (PCRs) with all mentioned methods perform similarly to PCR using dummy coding with autoscaling. When the proportion of non-metric variables is low, PLS-based methods show relatively small differences. With a high proportion of non-metric variables PLSR with the aggregation method, optimal scaling method, NM-PLSR and normal mean method show larger deterioration than other PLS-based methods. Ordinal PLSR is the worst PLS-based method when the expected number of categories is high.

In general, dummy PLSR with autoscaling performs best in all settings. Furthermore, dummy coding is easy to implement and interpret. Therefore, we focus on dummy coding in the following sections.

1.3 Applications

In this section we consider three applications. The first two applications generate wealth indices with two different responses and the third one uses the KOF Index of Globalization to predict economic growth.

1.3.1 Data

The first data set is the Demographic Health Survey (DHS, Central Bureau of Statistics (CBS) Kenya et al., 2004) from Kenya 2003. DHS is a widely used survey instrument to generate data on population, health and nutrition. Since the survey does not include incomes, a wealth index is commonly used as a proxy for socioeconomic status. The variables used to construct the wealth index describe possession of consumer durables, the type of housing and access to services that are selected and coded following Rutstein and Johnson (2004). There are in total 1 metric and 14 categorical variables, 10 of which are binary. The Body Mass Index (BMI) for the adult population is taken as an outcome variable, which is expected to be affected by household wealth (Wittenberg, 2013). A low BMI points to problems of serious undernutrition which is substantial in Kenya, while a high BMI points to overweight, which is also an emerging problem in the country (Rischke et al., 2014). But it is not clear that the weights for the wealth index arrived at by PCA will be the best predictor of the BMI, so that comparing the results with PLS is instructive. The data set has complete observations on 6686 individuals.

The second data set is the Indonesian Family Life Survey (Strauss et al., 2004) from the year 2000. Variables are selected and coded similarly to the DHS data. There are 11 categorical variables, with 8 of them being binary. As a dependent variable we consider log real monthly household expenditure per capita. We do this to investigate which weights best predict expenditures. A wealth index is often used to proxy for expenditures in many applications (where expenditures are not available) and thus the choice of appropriate weights is an important question. There are 10222 complete observations of households.

The third data set is from Dreher et al. (2008).[2] It consists of panel data with 23 metric variables capturing various facets of globalization. As an outcome variable, we focus on economic growth, which is expected to increase with globalization. Economic growth

[2] We use the 2013 version of the KOF index.

is measured as the annual growth rate of GDP per working age population. Since the KOF Index is an 'all-purpose' index of globalization, it is again instructive to study how the weights change if we condition them on a particular outcome variable. Clearly growth is determined not only by globalization, but also by other variables. Therefore, we include control variables following Bergh and Karlsson (2010) and Mankiw et al. (1992). Our control variables are initial GDP per working age population (Y0), a country's investment as a share of GDP (INV), the growth rate of the average years of schooling in the population (DHUM) and the growth rate of the working age population (DWAP). Growth and the control variables are constructed using data from Feenstra et al. (2013), the World Bank (2013) and Barro and Lee (2013). To smooth growth over the business cycle, we take 4 year averages of all variables.[3] We drop oil producing countries and countries where data quality is low (indicated as D grade in Feenstra et al. (2013)), as we suspected high measurement errors there. There are 575 complete observations including 63 countries and 10 time periods.

In our analysis we report the weights in both composite indices (PLS- and PCA-based) u_1 and w_1 and the corresponding regression coefficients $\widehat{\beta}_{PCR}$ and $\widehat{\beta}_{PLSR}$. More specifically, we proceed as follows. In the wealth index applications, all non-metric variables are transformed using dummy coding and afterwards autoscaling is applied, that is we work with $X_d^* = X_d D^{-1/2}$, where $X_d \in \mathbb{R}^{N \times k_d}$ contains metric variables and the indicator matrices from non-metric variables and $D = \text{diag}[\text{var}(x_{d,1}), ..., \text{var}(x_{d,k_d})]$ with $x_{d,j}$ denoting the j-th column of X_d. The weights u_1^* and w_1^* are derived from X_d^* and Y and the least squares estimator is obtained for Y, which can be expressed in terms of X_d. For example, for PLSR we obtain

$$\widehat{Y} = \widehat{\gamma}_0 + S\widehat{\gamma}_1 = \widehat{\gamma}_0 + X_d D^{-1/2} w_1^* \widehat{\gamma}_1 = \widehat{\gamma}_0 + X_d \widehat{\beta}_{PLSR}$$

[3] We use the geometric mean for growth rate variables and the arithmetic mean otherwise.

Hence, the reported PCR and PLSR regression coefficients are given in terms of X_d for the ease of interpretation. Analogously, weights are reported in terms of X_d, $u_1 = D^{-1/2}u_1^*$ and $w_1 = D^{-1/2}w_1^*$. Note that usually we cannot interpret $\widehat{\beta}_{PCR}$ and $\widehat{\beta}_{PLSR}$ as causal determinants, but rather aim to learn which variables are important predictors to the regressand.

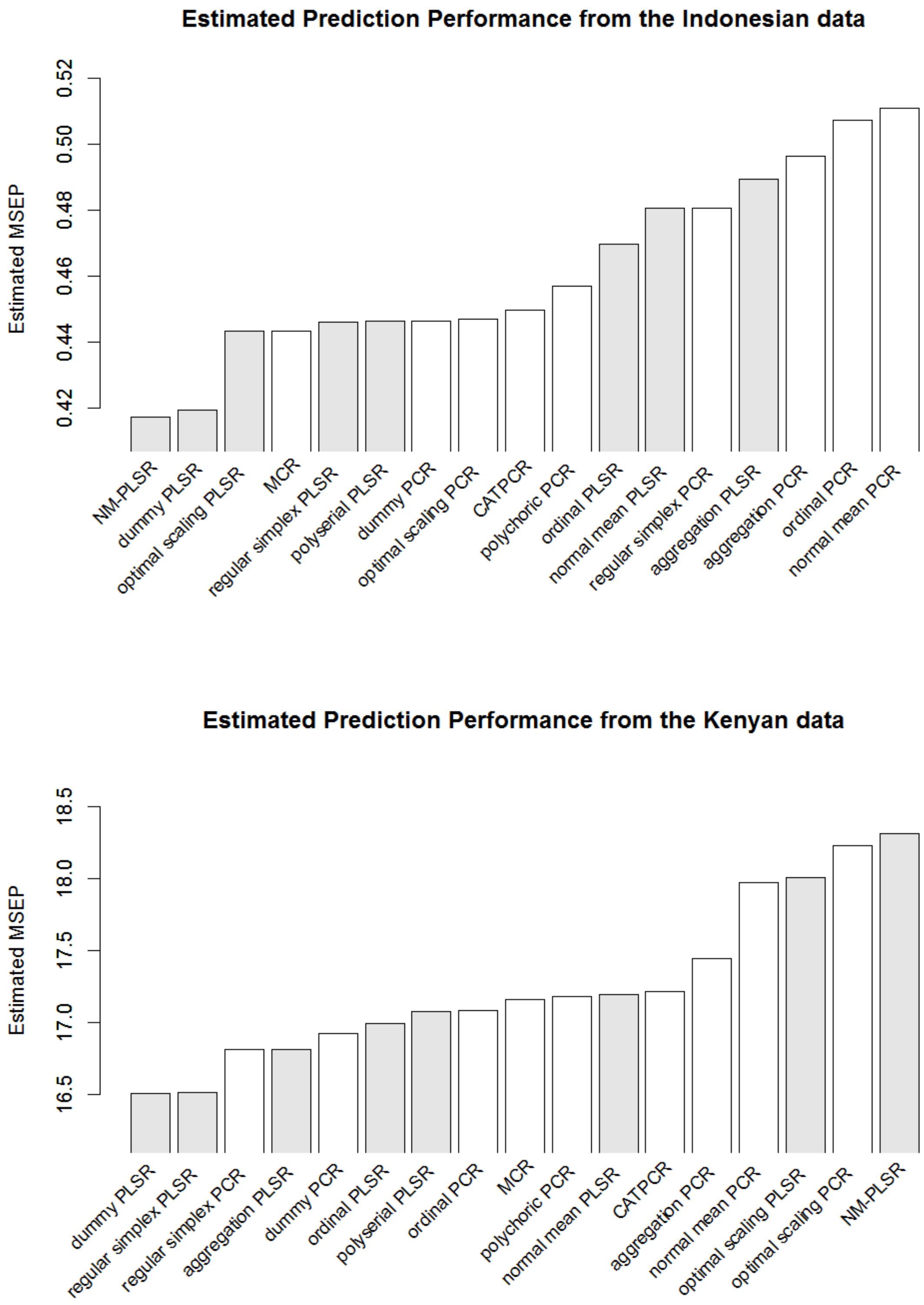
In the globalization application there are no non-metric variables and all the variables from Dreher et al. (2008) are already scaled for PCA or PLS. Therefore, no additional scaling is applied and $D = \text{diag}(1, 1, ..., 1)$.

Figure 1.1 shows the estimated prediction performance of various treatments on non-metric variables in PLS and PCA via 10-fold cross-validation (Mevik and Cederkvist, 2004) from the Indonesian and Kenyan applications. In analogy to the simulation study, PLSR using dummy coding performs excellently. It performs second best for the Indonesian data and best for the Kenyan data.

1.3.2 Wealth Index with BMI as the Outcome Variable

Table 1.2 shows the regression coefficients as well as the weights using PCA (left column) and PLS (right columns). The Jackknife standard errors (Martens and Martens, 2000) were used. The R^2 and the estimated MSEP for PLS are moderately better than for PCA (which is to be expected given that the correlation with the dependent variable is considered when creating the weights). More interesting are the differences in the weights. While the weights are quite similar for many indicator variables, they have the opposite sign in the case of bicycle and piped water at a public standpipe, suggesting that in order to predict the BMI, having a bicycle and access to a public standpipe both positively influence wealth. In quite a few variables, the size of the weights (while going in the same direction) differs substantially in magnitude. For example, using PLS, roofing is generally a more important driver of wealth (when predicting BMI), as is water access.

Figure 1.1: Estimated prediction performance of the various treatments of non-metric variables



PCA-based methods are colored white and PLS-based methods light grey. The MSEP is estimated via 10-fold cross-validation.

Table 1.2: PLS and PCA weights and the regressions with the outcome variable BMI in Kenya

	PCA			PLS		
	$\widehat{\beta}_{PCR}$	(se)	u_1	$\widehat{\beta}_{PLSR}$	(se)	w_1
electricity	0.428***	(0.018)	0.753	0.438***	(0.020)	0.680
radio	0.188***	(0.012)	0.331	0.313***	(0.020)	0.486
television	0.369***	(0.016)	0.649	0.426***	(0.019)	0.662
refrigerators	0.524***	(0.022)	0.921	0.434***	(0.033)	0.673
bicycle	−0.021***	(0.007)	−0.037	0.035*	(0.019)	0.054
motorcycle	0.193***	(0.045)	0.340	0.320**	(0.144)	0.496
car	0.443***	(0.021)	0.780	0.384***	(0.033)	0.595
telephone	0.424***	(0.017)	0.746	0.445***	(0.022)	0.690
servant	0.467***	(0.027)	0.821	0.307***	(0.039)	0.477
farm land	−0.160***	(0.009)	−0.282	−0.151***	(0.018)	−0.234
# hh member per room	−0.043***	(0.003)	−0.076	−0.083***	(0.005)	−0.129
water: piped in res.	0.355***	(0.016)	0.624	0.364***	(0.019)	0.565
water: piped public	−0.022***	(0.007)	−0.039	0.079***	(0.029)	0.122
water: inside well	0.002	(0.009)	0.003	0.011	(0.033)	0.018
water: surface	−0.235***	(0.012)	−0.414	−0.294***	(0.016)	−0.456
water: rain	0.015	(0.015)	0.026	0.255***	(0.063)	0.395
water: well public	−0.129***	(0.010)	−0.227	−0.150***	(0.026)	−0.233
toilet: own flush	0.505***	(0.020)	0.889	0.382***	(0.026)	0.592
toilet: shared flush	0.225***	(0.022)	0.395	0.261***	(0.043)	0.404
toilet: v.p. latrine	0.071***	(0.012)	0.126	0.202***	(0.037)	0.314
toilet: field	−0.248***	(0.016)	−0.436	−0.490***	(0.023)	−0.760
floor: dirt	−0.341***	(0.016)	−0.600	−0.409***	(0.017)	−0.635
floor: wood	0.378***	(0.069)	0.666	0.131	(0.101)	0.203
floor: cement	0.237***	(0.016)	0.417	0.359***	(0.019)	0.557
floor: tile	0.472***	(0.028)	0.830	0.289***	(0.043)	0.449
roof: natur	−0.257***	(0.016)	−0.451	−0.424***	(0.020)	−0.659
roof: iron	0.022*	(0.013)	0.039	0.227***	(0.020)	0.352
roof: tile	0.490***	(0.022)	0.861	0.366***	(0.032)	0.567
R^2	0.112			0.135		
$\widehat{\text{MSEP}}$	16.905			16.523		

Note: *** p<0.01, ** p<0.05, * p<0.1, As base categories “water: other”, “toilet: other”, “floor: other” and “roof: other” are excluded.

The differences in the weights transfer to the differences in the coefficients as well. For example, having a bicycle and access to a public standpipe predicts a low BMI in the PCR, whereas in the PLSR the prediction goes in the opposite direction. Roofing and water access are generally stronger predictors of BMI in the PLSR than the PCR.

Table 1.3: Correlations and prediction performance of PLS- and PCA-based wealth index with respect to socio-economic variables for the Kenyan data

		$\hat{\theta}_{pca}$	$\hat{\theta}_{pls}$	$\hat{\theta}_{pca} - \hat{\theta}_{pls}$ BS CI 95%
correlation	household size	-0.1829	-0.2185	[0.0330; 0.0381]
	# dead children	-0.1782	-0.1852	[0.0047; 0.0093]
	immunization (polyserial)	-0.0707	-0.0923	[0.0181; 0.0252]
MSEP	household size	7.0895	6.9848	[0.0959; 0.1141]
	# dead children	0.8867	0.8844	[0.0015; 0.0032]
	immunization (logit)	0.2119	0.2115	[0.0003; 0.0005]

Note: Individual data with N=31282. Bootstrapping percentile confidence interval with 10000 iterations.

In Table 1.3 we show that the wealth index created using PLS (with BMI as the outcome variable) also has a closer correlation to related health issues, such as whether child deaths occurred in the household, children are immunized, and household size. We check the prediction performance of the wealth indices to each variable using a simple linear regression, with an appropriate link function added if necessary. The prediction performance is again measured in terms of the estimated MSEP via 10-fold cross-validation. It appears that conditioning the weights for the wealth index on the correlation with a health-related outcome variable improves the predictive performance of the wealth index for other socio-economic outcomes.

Table 1.4: PLS and PCA weights and the regressions with outcome variable log household expenditure in Indonesia

	PCA			PLS		
	$\widehat{\beta}_{PCR}$	(se)	u_1	$\widehat{\beta}_{PLSR}$	(se)	w_1
electricity	0.168***	(0.006)	0.915	0.133***	(0.007)	0.629
television	0.112***	(0.003)	0.612	0.120***	(0.004)	0.568
refrigerators	0.149***	(0.006)	0.812	0.228***	(0.007)	1.081
vehicle	0.059***	(0.003)	0.323	0.054***	(0.004)	0.256
own: house	−0.065***	(0.003)	−0.357	−0.090***	(0.005)	−0.425
own: buildings	0.078***	(0.005)	0.426	0.116***	(0.008)	0.551
own: non-farm land	0.004	(0.004)	0.023	0.029***	(0.006)	0.137
own: farm land	−0.088***	(0.003)	−0.479	−0.045***	(0.005)	−0.215
water: piped	0.105***	(0.004)	0.571	0.091***	(0.005)	0.431
water: well	−0.047***	(0.004)	−0.257	−0.066***	(0.005)	−0.314
water: surface	−0.130***	(0.007)	−0.708	−0.096***	(0.008)	−0.455
water: rain	−0.045***	(0.017)	−0.248	−0.029	(0.021)	−0.139
water: basin	−0.090***	(0.016)	−0.493	−0.068***	(0.018)	−0.321
water: mineral	0.100***	(0.011)	0.547	0.248***	(0.020)	1.177
toilet: septank	0.136***	(0.003)	0.743	0.150***	(0.004)	0.713
toilet: no septank	−0.069***	(0.004)	−0.374	−0.054***	(0.006)	−0.257
toilet: communal	−0.019***	(0.005)	−0.103	−0.004	(0.009)	−0.019
toilet: public	−0.009*	(0.006)	−0.050	−0.054***	(0.011)	−0.257
toilet: field	−0.124***	(0.004)	−0.677	−0.150***	(0.005)	−0.708
cooking: electricity	0.035**	(0.015)	0.190	0.200***	(0.045)	0.948
cooking: gas	0.134***	(0.007)	0.732	0.228***	(0.008)	1.079
cooking: kerosene	0.076***	(0.003)	0.413	0.019***	(0.004)	0.092
cooking: wood, coal	−0.154***	(0.003)	−0.838	−0.163***	(0.004)	−0.772
cooking: don't cook	0.041***	(0.007)	0.223	0.247***	(0.021)	1.172
R^2	0.211			0.260		
$\widehat{\text{MSEP}}$	0.446			0.419		

Note: *** p<0.01, ** p<0.05, * p<0.1, As base categories "water: other", "toilet: other" and "cooking: other" are excluded.

1.3.3 Wealth Index with Expenditure as the Outcome Variable

In Table 1.4, we show the weights using PCA and PLS with expenditures as the outcome variable using our Indonesian data set. As the wealth index is often used as a proxy for expenditures, using PLS seems particularly appropriate to derive the weights for such a wealth index. Several features are noteworthy. First, the R^2 is somewhat improved using PLS, more so than in our first application suggesting that much new information is gained when the correlation with the outcome variable is considered. The PLSR again outperforms the PCR in terms of the estimated MSEP. Clearly when one wants to use the wealth index as a proxy for expenditures, it would be better to use the weights generated by PLS. Second, while the signs of the weights do not differ between PLS and PCA, the size of the weights differs substantially. For example, cooking materials and ownership of a fridge is generally more important in the PLS, electricity seems to be less important. In analogy to the weights, the PLSR and PCR coefficients show large differences. In the PLSR owning non-farm land predicts large household expenditure and using a public toilet predicts small household expenditure, whereas the PCR neglects them. Using rainwater as drinking water and using a communal toilet are not important predictors in the PLSR, but the PCR finds them to be significant. Cooking material and refrigerators are generally strong predictors, while electricity less strong predictor in the PLSR compared to the PCR.

Table 1.5 shows that using the PLS wealth index also generates slightly improved correlations with socio-economic outcomes such as school attendance or days sick. Additionally, the PLS wealth index predicts those variables slightly better.

1.3.4 Globalization Index with Growth as the Outcome Variable

Table 1.6 shows the results for the first stage regression, where we explain growth with its initial level (Y0), investment (INV), human capital (DHUM), population growth (DWAP)

Table 1.5: Correlations and prediction performance of PLS- and PCA-based wealth index with respect to socio-economic variables for the Indonesian data

		$\hat{\theta}_{pca}$	$\hat{\theta}_{pls}$	$\hat{\theta}_{pca} - \hat{\theta}_{pls}$ BS CI 95%
correlation	ever attended school (polyserial)	0.0496	0.0607	[-0.0158 ; -0.0065]
	# days being sick last month	-0.0219	-0.0288	[0.0035; 0.0104]
MSEP	ever attended school (logit)	0.2363	0.2362	[0.0001; 0.0003]
	# days being sick last month	1.9413	1.9407	[0.0002; 0.0013]

Note: Individual child data with N=11668. Bootstraping percentile confidence interval with 10000 iterations.

Table 1.6: The first stage regression

	$\widehat{\text{coef}}$	(se)
Y0	-0.598^{***}	(0.210)
INV	0.075^{***}	(0.027)
DHUM	-0.157	(0.097)
DWAP	0.147	(0.234)
R^2	0.137	

Note: Country fixed effects are included. *** p<0.01, ** p<0.05, * p<0.1

and country fixed effects. The results are in line with the previous literature (e.g. Mankiw et al., 1992). They show conditional convergence, at the one percent level of significance. Also at the one percent level, growth increases with investment, while human capital and population growth are not significant at conventional levels. We use the residuals from the regression as the outcome variable for comparing the effect of globalization on growth using PLSR and PCR, respectively, thereby holding these standard covariates constant. In other words, we compare the effect of globalization on those parts of economic growth that are not explained by its conventional determinants.

Both of the resulting indices (i.e. using PLS and PCA respectively) have positive and significant effects on growth when these covariates were controlled for, a result which is in line with the existing literature (e.g. Dreher, 2006; Rao et al., 2011). The result is not reported, but available upon request.

Table 1.7: PLS and PCA weights and the regressions with outcome variable growth

	PCA			PLS		
	$\widehat{\beta}_{PCR} \times 10^6$	(se$\times 10^6$)	u	$\widehat{\beta}_{PLSR} \times 10^6$	(se$\times 10^6$)	w
trade	6.077***	(2.063)	0.160	6.543	(6.319)	0.093
FDI	7.436***	(2.537)	0.196	25.611***	(7.233)	0.366
portfolio inv.	6.271***	(2.150)	0.165	12.507**	(5.785)	0.179
pay. foreigners.	6.805***	(2.299)	0.180	12.422	(7.641)	0.177
hidden import barriers	7.489***	(2.514)	0.198	2.377	(7.064)	0.034
tariff rate	10.619***	(3.482)	0.280	13.277*	(6.850)	0.190
taxes on trade	8.110***	(2.685)	0.214	1.726	(5.063)	0.025
CA restrict.	9.979***	(3.352)	0.263	20.001***	(6.424)	0.286
tele. traffic	9.021***	(2.993)	0.238	12.773***	(4.688)	0.182
transfers	1.901**	(0.813)	0.050	15.720**	(6.694)	0.225
tourism	8.142***	(2.704)	0.215	5.791	(5.064)	0.083
foreign pop.	7.397***	(2.404)	0.195	−0.695	(7.296)	−0.010
Int'l letters	5.801***	(1.974)	0.153	−4.401	(6.334)	−0.063
internet	9.129***	(3.076)	0.241	30.244***	(6.640)	0.432
television	6.134***	(2.020)	0.162	1.690	(4.247)	0.024
newspapers	7.548***	(2.536)	0.199	5.924	(6.196)	0.085
McDonald	12.429***	(4.156)	0.328	23.396***	(8.894)	0.334
Ikea	12.383***	(4.138)	0.327	7.563	(6.439)	0.108
books	5.471***	(1.867)	0.144	4.803	(5.508)	0.069
embassies	2.445***	(0.927)	0.065	5.715	(6.280)	0.082
Int'l org.	4.199***	(1.527)	0.111	22.767**	(8.924)	0.325
UNSC	10.895***	(3.690)	0.288	20.636**	(9.572)	0.295
Int'l treaties	5.103***	(1.782)	0.135	16.855*	(8.882)	0.241
R^2	0.012			0.029		
$\widehat{\text{MSEP}}$	0.000856			0.00085		

Note: *** p<0.01, ** p<0.05, * p<0.1, Dashed lines divide economic, social and political globalization.

We turn to our disaggregate analysis in Table 1.7. As can be seen at the bottom of the table, the R^2 of the PLSR is larger, while the estimated MSEP (using the Jackknife) is slightly smaller, compared to those of the PCR. Overall, the PLS procedure gives weights and a corresponding score which lead to better fit and prediction than the PCA. The table also reports the coefficients of the components of the KOF index. As can be seen, the results are in line with the previous literature, with most coefficients showing positive and significant correlations with growth when determining the weights using PCA. The table also shows the weights we obtain for the individual components.[4] The results differ substantially when we use PLS rather than PCA (right column of Table 1.7). Almost half of the variables are no longer significant at conventional levels. It could be because PLS has consumed more degrees of freedom compared to PCA (see Krämer and Sugiyama, 2011). Regarding actual economic flows, we find that economic growth increases with a country's stock of FDI and portfolio investments (both in percent of GDP on the original scale[5]), but not with its trade volume (also in percent of GDP). With respect to restrictions, the absence of restrictions on the capital account and lower mean tariff rates associate with growth positively, at the one and ten percent level of significance, respectively, while hidden import barriers and taxes on trade are not significant at conventional levels.

Concerning social globalization, few of the 11 indicators are significant at conventional levels. Specifically, economic growth increases with the amount of international telephone traffic, transfers received and given without a quid pro quo, the number of internet users, and the number of McDonalds restaurants in a country (as an indicator of cultural globalization). Conversely, three out of four indicators of political globalization are positively correlated with growth: the number of international organizations the country is a member of, the participation in the United Nations Security Council missions, and the number

[4] Note that these weights differ from those of the original index, given that we apply the PCA to our particular sample.

[5] Note that the KOF indices transform the original data on a percentile scale, so that they range between 1 and 100, with higher values showing more globalization.

of treaties signed.

Table 1.8: Correlations and prediction performance of PLS- and PCA-based globalization index with respect to physical integrity and empowerment rights

		$\hat{\theta}_{pca}$	$\hat{\theta}_{pls}$	$\hat{\theta}_{pca} - \hat{\theta}_{pls}$ BS CI 95%
correlation	physical integrity (polyserial)	0.6988	0.5545	[0.1281; 0.1606]
	empowerment rights (polyserial)	0.5516	0.4993	[0.0334; 0.0714]
MSEP	physical integrity (ordered logistic)	4.1508	6.2692	[-2.8278; -1.4446]
	empowerment rights (ordered logistic)	9.2684	9.9715	[-1.1132; -0.1039]

Note: Cross-country panel data with N=1581. Bootstraping percentile confidence interval with 10000 iterations.

Table 1.8 shows the correlations and MSEPs of the PLS- and PCA-based globalization indices with respect to physical integrity and empowerment rights, taken from the Cingranelli-Richards Human Rights Dataset (CIRI; Cingranelli and Richards, 2006). According to the recent survey on consequences of globalization in Potrafke (2014), improvements in human rights are among the important correlates of globalization. We rely on two indices: Physical integrity rights measure the absence of torture, extrajudicial killings, political imprisonments, and disappearances, on a scale of 0-8. Empowerment rights comprise the freedom of movement, freedom of speech, workers' rights, political participation, and freedom of religion, ranging from 0-10. On both indices, higher values represent better human rights practices.

The results of Table 1.8 show that both the PLS- and the PCS-based indices are positively correlated with physical and empowerment rights, at the five percent level of significance. For both indices, the PCA-based index performs "better," showing higher correlations and lower MSEPs. Given that the weights for the PLS-based index have been constructed to explain growth rather than human rights, this is unsurprising. Still, the high correlation with an established correlate of globalization is reassuring.

1.4 Conclusions

In this paper, we use both PCA and PLS to generate composite indices. Various treatments of non-metric variables in PCA and PLS are compared by means of a simulation study and we find that PLS with dummy coding not only performs better than more sophisticated statistical procedures, but is also easy to implement and interpret. This finding also holds for the real data considered in this paper. In our applications, PLS generates different weights and coefficients from PCA, which lead to better prediction and model fit of PLSR compared to PCR. We have checked whether composite indices based on PLS have a higher correlation or better prediction performance to different outcome variables, which works for two out of our three applications. We argue that when using statistical procedures to generate composite indices, it is not clear that the methods currently most commonly used, i.e. those based on the correlation between the indicator variables, are superior to derive weights. Often it may be more appropriate to create composite indices with particular outcomes in mind and PLS is a useful way to do so.

1.A Descriptions of Variables

Table 1.9: Variable names and variable labels of the Kenyan data

variable names	variable labels
electricity	electricity
radio	radio
television	television
refrigerators	refrigerators
bicycle	bicycle
motorcycle	motorcycle
car	car
telephone	telephone
servant	domestic servant
farm land	own farm land
# hh member per room	number of household members per room
water: piped in res.	piped water in residence
water: piped public	piped water in public
water: inside well	inside well water
water: surface	surface water
water: rain	rain water
water: well public	public well water
toilet: own flush	own flush toilet
toilet: shared flush	shared flush toilet
toilet: v.p. latrine	ventilated pit latrine toilet
toilet: field	bush field toilet
floor: dirt	dirt floor
floor: wood	wood floor
floor: cement	cement floor
floor: tile	tile floor
roof: natur	natural roof
roof: iron	iron roof
roof: tile	tile roof

Table 1.10: Variable names and variable labels of the Indonesian data

variable names	variable labels
electricity	electricity
television	television
refrigerators	refrigerators
vehicle	vehicle
own: house	own house
own: buildings	own other buildings
own: non-farm land	own non-farm land
own: farm land	own farm land
water: piped	piped water
water: well	well water
water: surface	surface water
water: rain	rain water
water: basin	basin water
water: mineral	mineral water
toilet: septank	toilet with septic tank
toilet: no septank	toilet without septic tank
toilet: communal	communal toilet
toilet: public	public toilet
toilet: field	field toilet
cooking: electricity	electricity cooking
cooking: gas	gas cooking
cooking: kerosene	kerosene cooking
cooking: wood, coal	wood or coal cooking
cooking: don't cook	don't cook

Table 1.11: Variable names and variable labels of the globalization data

variable names	variable labels
trade	Trade (percent of GDP)
FDI	Foreign Direct Investment, stocks (percent of GDP)
portfolio inv.	Portfolio Investment (percent of GDP)
pay. foreigners.	Income Payments to Foreign Nationals (percent of GDP)
hid. im. barriers	Hidden Import Barriers
tariff rate	Mean Tariff Rate
taxes on trade	Taxes on International Trade (percent of current revenue)
CA restrict.	Capital Account Restrictions
tele. traffic	Telephone Traffic
transfers	Transfers (percent of GDP)
tourism	International Tourism
foreign pop.	Foreign Population (percent of total population)
Int'l letters	International letters (per capita)
internet	Internet Users (per 1000 people)
television	Television (per 1000 people)
newspapers	Trade in Newspapers (percent of GDP)
McDonald	Number of McDonald's Restaurants (per capita)
Ikea	Number of IKEA (per capita)
books	Trade in books (percent of GDP)
embassies	Embassies in Country
Int'l Org.	Membership in International Organizations
UNSC	Participation in U.N. Security Council Missions
Int'l treaties	International Treaties

Chapter 2

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

with Tatyana Krivobokova

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 an extensive simulation study under several typical data generating processes and associated recommendations are made. PLS-based methods are to prefer in practice, since, independent of data generating process, PLS performs either as good as PCA or significantly outperforms it. PLS with dummy coding and NM-PLSR are often prefered treatments of non-metric variables. An application of PLS and PCA algorithms with non-metric variables is considered, which generates wealth indices to pre-

dict household expenditures. In analogy to the simulation studies, PLS outperforms PCA, and model selection statistics support dummy coding.

2.1 Introduction

Principal Component Analysis (PCA, Hotelling, 1933) and Partial Least Squares (PLS, Wold, 1966b) 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. This simulation study differs from the simulation study in Chapter 1 in that a latent variable of interest is explicitly assumed. PLS with dummy coding shows generally good performance, especially when the latent variable of interest account for only small variations in the regressor matrix.

As an application, we consider construction of wealth indices 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 2.2 recapitulates PCA and PLS algorithms and reviews the treatments of non-metric variables in PCA and PLS in the literature. In Section 2.3 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 2.4, before we conclude in Section 2.5.

2.2 PCA and PLS with Non-metric variables

2.2.1 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.

2.2.2 Treatments of Non-metric Variables in PCA and PLS

Treatments of non-metric variables in PCA and 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 $\tilde{x}_{ij} = \mathrm{Ver}_{m_j - 1}(x_{ij}) \in \mathbb{R}^{1 \times m_j}$, where $\mathrm{Ver}_{m_j - 1}(x_{ij})$ 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 scaled, 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, $j = 1, \dots, K_n$. 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, $w_1 = \text{Cor}(y, X)/\left\|\text{Cor}(y, X)\right\|$, where $\text{Cor}(y, X)$ 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).

2.3 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.

2.3.1 Simulation Design

We adapt the simulation designs from Naes and Martens (1985) and Kolenikov and Angeles (2009) with some adjustments. All simulation designs rely 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 in other observable variables. For example, in a wealth index application, one cannot observe household wealth directly, but wealth is assumed to be manifested in 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, the 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 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

$$x^*_{ij} = \Xi_{i1}\lambda_{1j} + \Delta_{ij}, \ \ i = 1, \dots, N, \ \ j = 1, \dots, K.$$

Here $\lambda_{1j} = 1/\sqrt{K}$, $j = 1, \dots, K$ 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 $\Delta_i \sim \mathcal{N}_K(0_K, I_K/(9K))$, such that the signal to noise ratio $\sqrt{\sum_{j=1}^{K} \text{Var}(\Xi_{i1}\lambda_{1j})/\sum_{j=1}^{K} \text{Var}(\Delta_{ij})} = 3$. 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

$$y_i = \Xi_{i1}\beta_1 + \varepsilon_i, \ \ i = 1, \dots, N, \tag{2.1}$$

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., (2.1) 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 $D_i \sim \text{Bin}(1, 0.5)$, $\beta_2 = \beta_3 = 1$ 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.

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

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 in all DGPs 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 each 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 (e.g., Dreher, 2006; Filmer and Pritchett, 2001; Rutstein and Johnson, 2004).

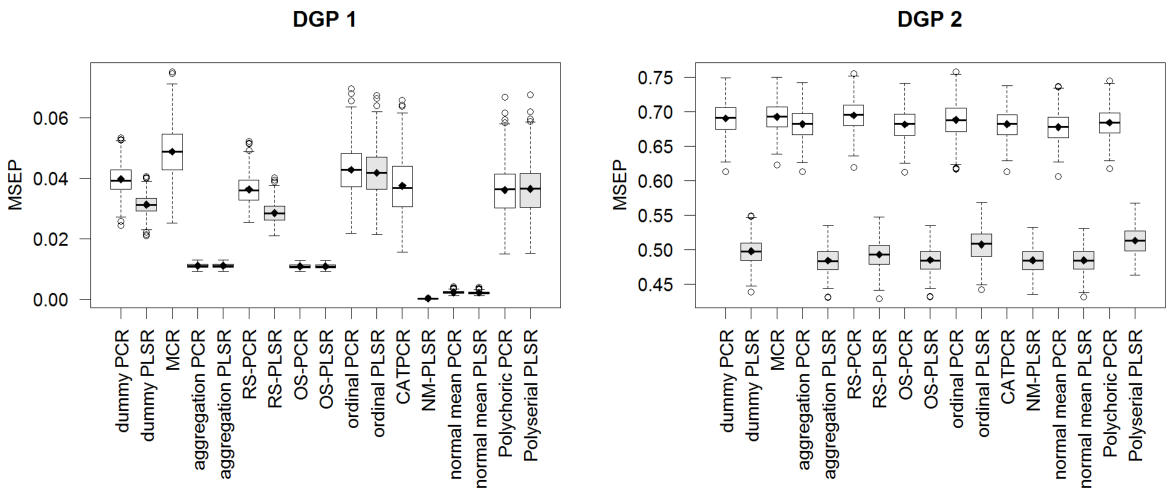
2.3.2 Simulation Results

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 only with nominal quantification are reported.

Figure 2.1 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 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

Figure 2.1: MSEP under DGP 1 (left) and DGP 2 (right)

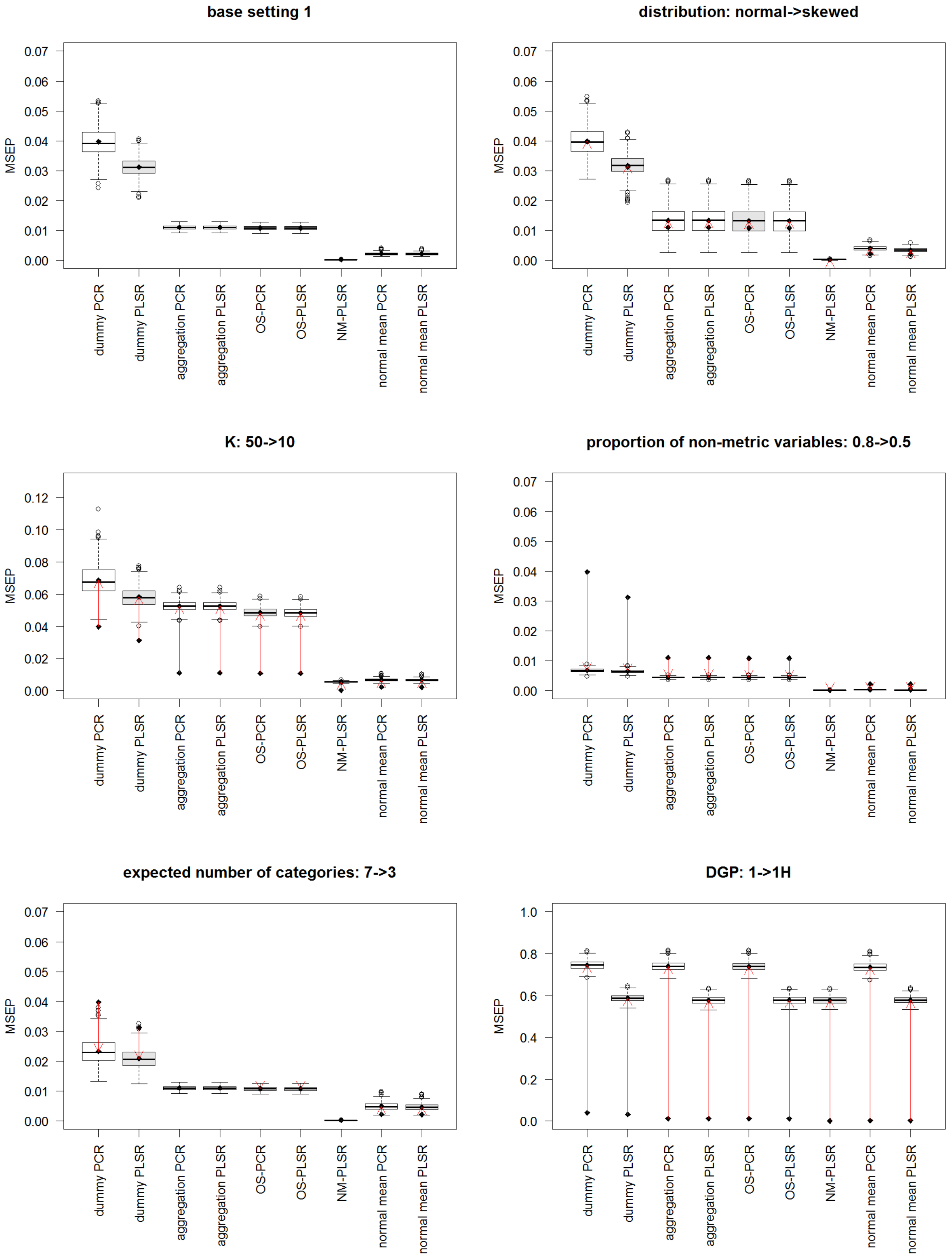


Base setting 1 and 2 are reported. PCA-based methods are colored white and PLS-based methods light grey.

more detail in Figure 2.2.

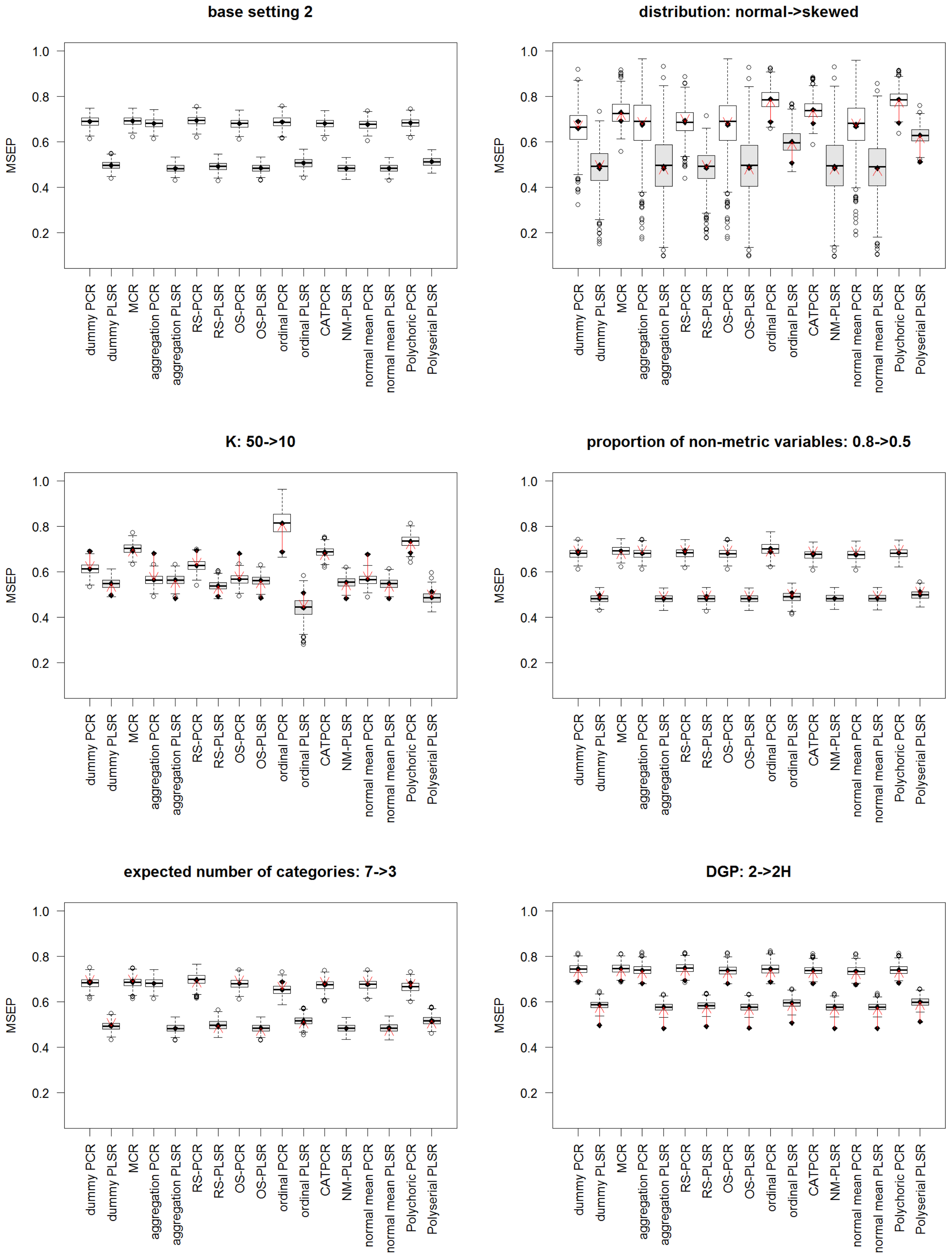
Figure 2.2 shows the performance of various methods under DGP 1. Note that scales on the left middle and right bottom plots are different from the other plots. We focus on *Base setting 1*, shown again in the left top plot and vary one setting at each subsequent plot. The changes of the means from the 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 2.1. 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

Figure 2.2: MSEP under DGP 1



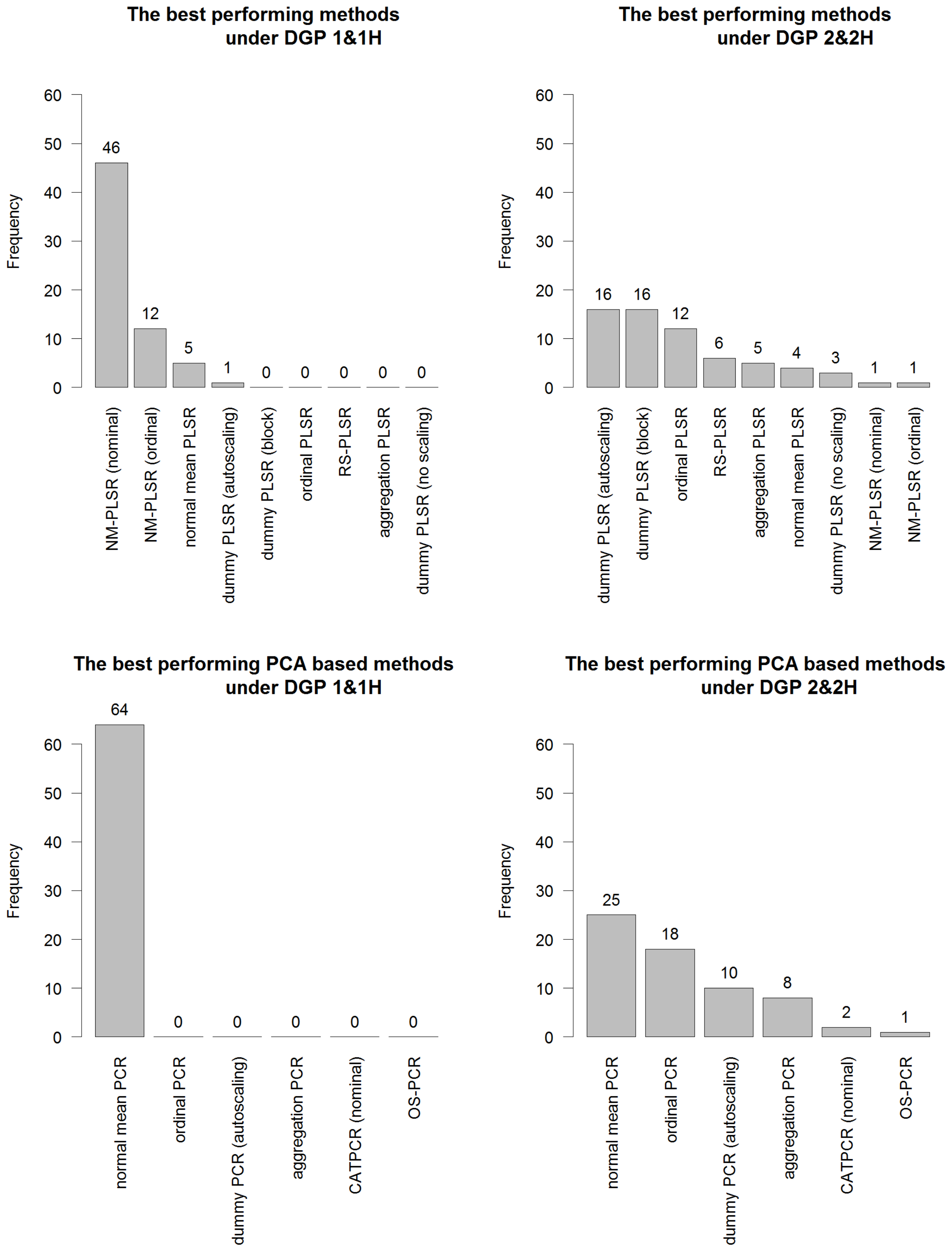
Base setting 1 is used. Red arrows mark changes of the means from the base setting to the respective setting. PCA-based methods are colored white and PLS-based methods light grey.

Figure 2.3: 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 2.4: The absolute frequency of the best perfoming methods over different DGP



improve with less expected number of categories.

The upper left panel of Figure 2.4 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. While some methods are not reported in Figure 2.2 to make the comparison easier, all methods are considered in Figure 2.4. 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 2.3 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 2.4 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, 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. Second, ordinal PCR significantly outperforms dummy PCR in 20 settings, which typically have normal distributed latent variable.

2.4 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:

$$y_i = \underline{1}_i\gamma_0 + T_i\gamma_1 + D_i\gamma_2 + T_i \otimes D_i\gamma_3 + \varepsilon_i,$$

where $\underline{1}_i$ is the intercept, 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_0, 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 2.2.2, 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 2.1 shows the coefficient estimates $\hat{\gamma}_1$ for PCA or PLS and model selection statistics. We show not only our favored model, i.e., the model with two scores with the control variables, but also models without the second score or the control variables. In this way we can see which parts of the model contribute to the performance. The PLSRs show better performance than the PCRs in terms of R^2 and estimated MSEP. In other words, composite indices based on PLS is better than PCA-based ones in terms of fitting and prediction in our application. Adding an additional score and controlling heterogeneity bring gains in terms of R^2 and estimated MSEP to both PLSRs and PCRs, whereby the

Table 2.1: Coefficient estimates in terms of composite indices and model selection criteria

	$\hat{\gamma}_{1,PCR}$ $A=1$	$\hat{\gamma}_{1,PCR}$ $A=1,H$	$\hat{\gamma}_{1,PCR}$ $A=2$	$\hat{\gamma}_{1,PCR}$ $A=2,H$
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
	$\hat{\gamma}_{1,PLSR}$ $A=1$	$\hat{\gamma}_{1,PLSR}$ $A=1,H$	$\hat{\gamma}_{1,PLSR}$ $A=2$	$\hat{\gamma}_{1,PLSR}$ $A=2,H$
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.

magnitude of the gains is larger by the PLSRs. The larger gains of the PLSRs compared to the PCRs could be bacause of the quality of the second score, which are significant in the PLSRs, but not in the PCRs. On the other hand the first scores are significant in all regressions. Wealth, measured by the first and second PLS scores, predicts higher houshold expenditures in the PLSR under the favored setting. The interpretation of the PCR is analogous, except that the second score is not significant. The inference was based on the Jackknife standard errors (Martens and Martens, 2000).

Table 2.2 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 2.1. 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

Table 2.2: 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.

compared to the PCR. Analogously, PLS and PCA weights show strong differences, with the PLS 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. A comparison of coefficient estimates between one and two scores model shows the contribution of the elements in the second score in the final prediction. Introducing the second score brings larger changes in coefficient estimates in the PLSR compared to the PCR, which is not surprising given that the PCR coefficient estimate in terms of the second score in Table 2.1 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.

2.5 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. PLS is particularly advantageous when informative variations account for small variances in the variables in a composite index (DGP 2&2H). When informative variations account for large variances (DGP 1&1H), PLS performs as good as PCA. 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 among PCA based methods, 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.

Chapter 3

An Application of Partial Least Squares to the Construction of the Social Institutions and Gender Index (SIGI) and the Corruption Perception Index (CPI)

with Stephan Klasen

Abstract

In this paper the Social Institutions and Gender Index (SIGI) is re-constructed using weights generated by Principal Component Analysis (PCA) and Partial Least Squares (PLS). Using the revised SIGI, we test the effects of social institutions related to gender inequality on several development outcomes, such as female education, fertility and child *mortality, controlling for relevant determinants. We also use the same procedure to study the relationship between the SIGI and corruption measued by the Corruption Percetion Index (CPI). Also for the CPI we consider alternative weighting procedures using PCA and PLS. We find that gender inequality in social institutions has significant effect on fertility and corruption regardless of the weighting procedure, while for female education and child mortality only the SIGI based on PLS generates significant results.*

3.1 Introduction

Gender inequlity not only deprives the women of basic freedom, but also hinders the development of the society, e.g., it has been found to cause ill-health, low overall human capital, bad governance, and lower economic growth (Branisa et al., 2013; Sen, 1999). This study focuses on the *social institutions related to gender inequality*, which shape societal practices and legal norms, ultimately producing gender inequality.

To measure a latent concept such as the social institutions related to gender inequality, a composite index is a natural approach. We build new composite indices besed on the indicators included in the *Social Institution and Gender Index* (SIGI; Branisa et al., 2013). The quality of a composite index depends on the weighting scheme. In Branisa et al. (2013) weights of the SIGI are derived as a mixture of polychoric principal component analysis (Kolenikov and Angeles, 2009) and the authors' judgement, which can be subjective. Therefore, we change the weighting scheme to Principal Component Analysis (PCA; Hotelling, 1933). However, PCA works when the largest variations in the variables building composite indices are informative, but in practice this is not always the case. We additionally use Partial Least Squares (PLS; Wold, 1966b) to derive weights, which considers the relationship between outcome variables and the variables building composite indices. Consequently, PLS often works well even when informative variations in the variables are small. When coefficient estimates from Principal Component Regression (PCR) are insignificant and Partial Least Squares Regression (PLSR) show significant coefficient estimates, we can suspect that PCA doesn't work well due to large noise. On the other hand, when both PCR and PLSR show insignificant coefficient estimates, we can be more sure about no relationship. Using PLS to derive weights has the following additional advantages. First, PLS usually builds composite indices better for the prediction of outcome variables compared to PCA when only few number of PCA or PLS scores are used (Naes and Martens, 1985). Second, a comparison between PCA and PLS weights shows which variables are particularly relevant for the prediction of a certain outcome variable.

The SIGI with new weights will be used to test the effects of social institutions related to gender inequality on various gender outcomes. In analogy to Branisa et al. (2013), we take *female education*, *fertility*, *child mortality* and *corruption* as the outcome variables. Branisa et al. (2013) found that the SIGI as a whole did not have an impact on these outcomes once control variables were included. They did, however, find that particular sub-indices of the SIGI had a significant impact on these outcome variables. We want to investigate here whether these results change if the SIGI is generated using PLS or PCA. In particular, we would like to investigate whether the reweighted SIGI as a whole has an impact on these outcomes. The weights of the SIGI that lead to such a significant relationship would then also yield new insights about the components of social institutions that are particularly relevant for different development outcomes. We perform a linear regression analysis for each outcome variable, while relevant control variables are added based on the literature. We check the non-linearity of the control variables and adjust the empirical model accordingly based on model selection criteria. Additionally, most indicators that are included in the SIGI are non-metric, for which special treatments are necessary to apply PCA and PLS. We compare various treatments for non-metric variables in terms of model selection criteria and choose dummy coding as the most appropriate treatment.

As we investigate the relationship between the SIGI and corruption, we use the *Corruption Perception Index* (CPI; Transparency International, 2013) as a measure of corruption. The CPI assigns weights via a simple average, which is appropriate when all variables are equally important, but it is not clear whether this condition is satisfied. One can suspect that many varaibles in the CPI have high measurement errors and some variables are emphasized without clear reasons. We modify the CPI by preparing variables differently and changing the weighting procedure to PCA and PLS and check the relationship between the SIGI and corruption again.

The rest of this paper is organized as follows. Section 3.2 recapitulates PCA and PLS algorithms. Section 3.3 discusses the data. Empirical analyses follow in Section 3.4. In Section 3.5, we create new CPIs with different weighting schemes. Then we conclude.

3.2 PCA and PLS Algorithms

We recapitulate PLS and PCA algorithms in the following. Consider a regression model $Y = X\beta + \varepsilon$, where $Y \in \mathbb{R}^{N \times R}$, $X \in \mathbb{R}^{N \times K}$, $\beta \in \mathbb{R}^{K}$, R, $K \leq N$ and $\varepsilon \in \mathbb{R}^{N}$ with $\mathbb{E}(\varepsilon|X) = 0$ and $cov(\varepsilon|X) = \sigma^2 I_n$. Note that outcome variables can be multivariate. In the following, we restrict our attention to the case where we have only a single interesting score from X or Y respectively. It is common in practice to assume the unidimensionality of a composite index, e.g., the KOF Index of Globalization (Dreher, 2006) and the wealth index (Rutstein and Johnson, 2004; Kolenikov and Angeles, 2009). Alternatively one can decide the number of scores based on model selection criteria (Wold et al., 1983; Zwick and Velicer, 1986), which is not pursued here.

Both PCA and PLS build the first score as a linear combination of the columns of regressor matrix and regressand matrix, that is $t_1 = Xw_1$ and $u_1 = Yc_1$. PCA builds the first score by maximizing the empirical variance of the score in terms of the weights.

$$w_1 = \underset{\|\omega_X\|=1}{\operatorname{argmax}} \, t_1' t_1 = \underset{\|\omega_X\|=1}{\operatorname{argmax}} \, \omega_X' X' X \omega_X$$

$$c_1 = \underset{\|\omega_Y\|=1}{\operatorname{argmax}} \, u_1' u_1 = \underset{\|\omega_Y\|=1}{\operatorname{argmax}} \, \omega_Y' Y' Y \omega_Y,$$

where t_1, $u_1 \in \mathbb{R}^N$, $w_1 \in \mathbb{R}^K$ and $c_1 \in \mathbb{R}^R$. The solution is the first eigenvector of X or Y respectively (Maitra and Yan, 2008). The first PLS score is identified by the maximization of the empirical covariance between the first score from X and Y.

$$\{w_1, c_1\} = \underset{\|\omega_X\|=\|\omega_Y\|=1}{\operatorname{argmax}} \, (t_1' u_1)^2 = \underset{\|\omega_X\|=\|\omega_Y\|=1}{\operatorname{argmax}} \, (\omega_X' X' Y \omega_Y)^2.$$

There are several algorithms to calculate the PLS weights (de Jong, 1993). In composite index applications weights are to be interpreted as the relative importance of the variables building a composite index.

3.3 Data

In this section we explain the variables that build the SIGI, our outcome variables and control variables. We take the concepts and data from Branisa et al. (2013) to build the SIGI. The SIGI is composed of 12 variables, which are divided into five blocks, and each block of variables builds a subindex. The subindices are generated by scaling the first polychoric PCA score (Kolenikov and Angeles, 2009) on domain $[0, 1]$. Then the subindices are squared and averaged to build the SIGI. The data cover about 100 non-OECD countries and the indicators are coded so that high value represents high gender inequality. The five blocs or dimensions of social institutions considered in the SIGI are **family code**, **civil liberties**, **physical integrity**, **son preference** and **ownership rights**. **Family code** is about the decision making power of women in the household, which is measured by the prevalence of early marriage (*Early marriage*), the prevalence of polygamy (*Polygamy*), whether women can become legal guardian of children or have custody right after divorce (*Parental authority*) and whether women have the rights to inherit (*Inheritance*). **Civil liberties** concern the freedom of social participation of women. They are measured by whether women can move outside freely without having to be escorted by men (*Freedom of movements*) and whether it is obligatory to wear a veil (*Freedom of dress*). **Physical integrity** refers to the violence against women, which is measured by the existence of legal protection for women againt rape, assault and sexual harrasment (*Violence against women*) and the prevalence of female genital mutilation (*Female genital mutilation*). **Son preference** measures the gender bias in mortality of girls compared with boys (*Son preference*), which is caused by sex selective abortion or inadequate care. **Ownership rights** cover the rights of women to several types properties. They are measured by the access to land (*Womens' access to land*), credit (*Womens' access to credit*) and properties other than land (*Womens' access to property other than land*). *Early marriage* and *female genital mutilation* are numerical variables and other indicators are ordinal variables.

We aim to test whether *female education*, *fertility*, *child mortality* and *corruption* are affected by the SIGI using the same hypotheses and measurements as Branisa et al. (2013). According to the hypotheses made in that paper, more gender inequality reduces female education, increases fertility, child mortality and corruption. *Female education* is measured by female gross secondary school enrollment rates (World Bank, 2008), which is the number of children in school divided by the population who are supposed to be in school by age in percent scale. *Fertility* is measured by total fertility rates (World Bank, 2009), which is the average number of birth to a woman in her lifetime. *Child mortality* is measured by child mortality rates (World Bank, 2008), that is under five mortality per 1000 live births. We take the Corruption Perception Index (CPI, Transparency International, 2013) as a measure of *corruption*, which is scaled from 0 to 10 with higher value indicating less corruption.

The control variables are taken from representative models from Branisa et al. (2013). All regressions control for the level of economic development, religion, region and the political system in a country. The level of economic development is measured by the log per capital GDP in constant price (*log GDP*, US$, PPP, base year 2005). Religion is measured by a Muslim majority dummy (*Muslim*) and a Christian majority dummy (*Christian*). Region dummies include East Asia and Pacific (*EAC*), South Asia (*SA*), Middle East and North Africa (*MENA*), Latin America and Caribean (*LAC*) and Europe and Central Asia (*ECA*). Sub-Saharan Africa (*SSA*) is the left out category. Political system is captured by the Electoral Democracy Index (*Electoral democ.*) and the Civil Liberties index (*FH civil liberties*) from Freedom House (2008), but for the corruption regression the Civil Liberties index is substituted by Polity 2 (*Polity 2*, Monty G. Marshall, 2013). The Civil Liberties index is coded in a way that high value means better analogous to other two variables. For the corruption regression, several additional control variables are added. Women's representation is controlled, which are measured by the proportion of female legislator (*Parliament*), the female share in professional, technical, admistrative and managerial positions (*Managers*) and women's share of labor force (*Labor force*), where all three variables are taken from World Bank (2008). We add ethnic fractionalization (*Ethnic frac.*, Alesina et al., 2003), literacy rates (*Literacy pop.*, United Nations Development Programme, 1995), trade openness (*Openness*, World Bank, 2008), a dummy indicating that a country has never been a colony and a British colony (*Not colony*, *British colony*, Correlates of War 2 Project., 2003).

Following Branisa et al. (2013), we take the average over five or six years (2000 or 2001-2005) for the regressands. The average over 10 years (1996-2005) is taken for the control variables.

We take the complete observations from a total of 124 of non-OECD countries for the regression analysis, which results in the number of observations for the female education regression as 91, the fertility regression as 97, the child mortality regression as 97 and the corruption regression as 85. We have checked whether there is a sample selection problem from the regressands regarding the dropped and kept observations by comparing the means using t-tests and the distributions using kernel density estimations and didn't find any suggestion of sample selection problem.

3.4 Empirical Analysis

Our empirical analysis proceeds with three steps. First, we formulate an empirical model. Second, we choose an appropriate treatments for non-metric variables in the SIGI when PCA or PLS are performed considering model selection statistics. We take the possible non-linearity between regressands and control variables into account during the selection. Third, we interpret the results from the selected models.

Our empirical analysis uses a simple linear model in analogy to Branisa et al. (2013).

$$u = \gamma_0 + SIGI\gamma_{SIGI} + Z\gamma_Z + \varepsilon,$$

where u is a regressand. The SIGI is the composite index and Z is a matrix containing control variables. γ_0, γ_{SIGI} and γ_Z are coefficient vectors of appropriate length and ε denotes an error term. We denote $\gamma_{PCR} = (\gamma_0, \gamma_{SIGI}, \gamma_Z)$ when the SIGI is calculated via PCA and γ_{PLSR} is analogously defined for the SIGI being calculated via PLS.

Next, we perform a model selection in terms of various treatments of non-metric variables for PCA and PLS available in the literature. The prediction performance measured by the estimated mean squared error of prediction (MSEP; Mevik and Cederkvist, 2004) via the Jackknife is considered as the model selection criterion. We focus on dummy coding with autoscaling because it performs usually good, albeit not always the best, and it is easy to implement and interpret compared with competing methods. The following methods are considered during the model selection, whereby a detailed summary of these methods is available in Chapter 2. Note that the abbreviation in the parenthesis corresponds to Figure 3.1. **Dummy coding** (dummy PCR/PLSR; Filmer and Pritchett, 2001), **multiple correspondence analysis** (MCR; Greenacre, 2010) and **regular simplex method** (RS-PCR/PLSR; Niitsuma and Okada, 2005) transform each unique category of a non-metric variable to a variable. **Optimal scaling method** (OS-PCR/PLSR; Tenenhaus and Young, 1985), **non-metric partial least squares regression** (NM-PLSR; Russolillo, 2009), **categorical principal component analysis** (CATPCR; Meulman, 2000) and **normal mean coding** (normal mean PCR/PLSR; Kolenikov and Angeles, 2009) scale each unique value of non-metric variables. **Polychoric PCR** (Kolenikov and Angeles, 2009) assumes that observed ordinal variables are generated from discretizations of multivariate normal latent variables. The variance-covariance matrix of the multivariate normal latent variables is estimated and used to calculate the weights of PCA. **Polyserial PLSR** is analogous to polychoric PCR, except that the weights are based on the polyserial correlation between outcome variable and ordinal variables. **Ordinal PCR** and **PLSR** consider ordinal variables as if they were numerical variables. The approach from Branisa et al. (**SIP.FGT;** 2013) as explained above is considered as a reference.

Next, we checked for non-linearity of control variables. The data suggested that log GDP has a non-linear effect on each outcome variable. We model the non-linearity by including linear, square and cubic term of log GDP, since more complicated non-parametric fits were not superior. In general, selected non-linear terms improved the estimated MSEP. The female education regression includes the linear term of log GDP, the fertility regression the linear and cubic terms, the child mortality regression the linear, square and cubic terms and the corruption regression the linear and cubic terms. In Figure 3.1, the performance of the various treatments in terms of the estimated MSEP under the selected non-linear terms are reported.

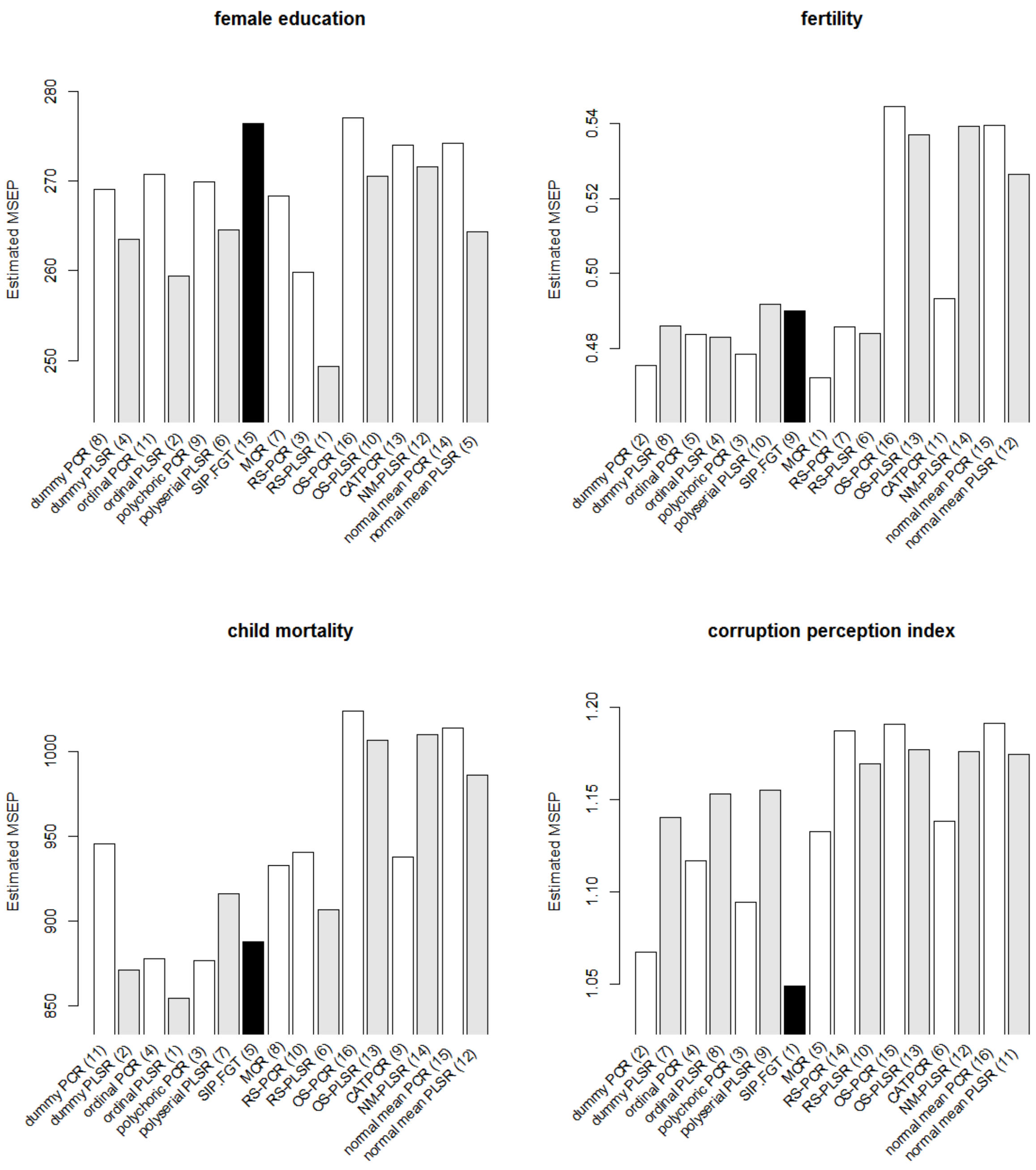
We report not only the coefficient estimates in terms of the SIGI, but also in terms of the variables building the SIGI. The coefficient estimates in terms of PCA or PLS score can be straightforwardly transformed back in terms of regressors.

$$\begin{aligned} u &= \hat{\gamma_0} + SIGI\hat{\gamma}_{SIGI} + Z\hat{\gamma}_Z + \hat{\varepsilon} \\ &= \hat{\gamma_0} + XS^{-\frac{1}{2}}w_1^*\hat{\gamma}_{SIGI} + Z\hat{\gamma}_Z + \hat{\varepsilon} \\ &= \hat{\gamma_0} + X\hat{\beta}_{SIGI} + Z\hat{\gamma}_Z + \hat{\varepsilon}, \end{aligned}$$

where X contains the variables building the SIGI after dummy coding and S is a scaling matrix, which is diagonal and containing the variance of each column of X. We report $\hat{\beta}_{PCR} = \hat{\beta}_{SIGI}$ and $w_{PCA} = S^{-\frac{1}{2}}w_1^*$ when the SIGI is calculated by PCA. When the PLS score is used for the SIGI, $\hat{\beta}_{PLSR}$ and w_{PLS} are reported, which are analogously defined.

Table 3.1 shows the results of the linear regressions for the outcome variables on the SIGIs built by PCA and PLS. The PLSRs fit data better than the PCRs for all outcome variables, which is visible through the higher R^2 of the PLSRs than the PCRs. The estimated MSEP of the PLSR is lower than the PCR for the female education and the child mortality regression, i.e., for those models PLS is beneficial to improve prediction. The inferences in the followings are based on the Jackknife standard errors (Martens and Martens, 2000). The SIGIs based on PCA have no significant effect on *female education* and *child mortality*, but the SIGIs based on PLS are significant at 5% and 1% level. It suggests that the weights generated by PCA generate SIGIs that have no significant impact on these outcomes, while the SIGIs generated by PLS have significant impact, where more gender inequality predicts lower female education and more child mortality. Considering PLS works often better than PCA when important latent variable has small variations in indicators, we can suspect large measurement errors are problemetic in the

Figure 3.1: Estimated MSEP of the various treatments for non-metric variables



MSEP is estimated via the Jackknife. PCA-based methods are colored white, PLS-based methods light grey and arbitrary methods black. Ascending ranks in the parenthesis.

Table 3.1: Linear regressions with the SIGI built by PCA and PLS

	female education		fertility		child mortality		CPI	
	$\hat{\gamma}_{PCR}$	$\hat{\gamma}_{PLSR}$	$\hat{\gamma}_{PCR}$	$\hat{\gamma}_{PLSR}$	$\hat{\gamma}_{PCR}$	$\hat{\gamma}_{PLSR}$	$\hat{\gamma}_{PCR}$	$\hat{\gamma}_{PLSR}$
SIGI	-2.65	-5.35^{**}	0.20^{**}	0.29^{***}	5.88	14.04^{***}	-0.23^{**}	-0.34^{**}
log GDP	12.60^{***}	10.73^{***}	-1.58^{***}	-1.40^{***}	-596.00^{*}	-561.42^{*}	-1.73^{*}	-1.98^{*}
(log GDP)^2					66.19^{*}	63.40		
(log GDP)^3			0.00^{**}	0.00^{*}	-2.49	-2.42	0.01^{**}	0.01^{**}
Muslim	1.33	3.28	0.39	0.33	26.62^{*}	19.79	0.05	-0.03
Christian	6.62	6.49	0.16	0.13	2.79	-0.07	-0.08	-0.05
SA	15.93^{*}	10.00	-1.74^{***}	-1.38^{***}	-58.08^{***}	-39.94^{**}	-0.18	-0.69
ECA	33.05^{***}	24.50^{***}	-1.88^{***}	-1.61^{***}	-66.04^{***}	-41.36^{**}	-0.88	-0.73
LAC	12.09	6.32	-0.44	-0.27	-50.30^{***}	-30.58^{**}	-0.70	-0.50
MENA	32.04^{***}	23.66^{**}	-1.32^{**}	-0.93^{*}	-95.93^{***}	-73.86^{***}	0.17	-0.22
EAP	18.27^{**}	10.35	-1.26^{***}	-0.99^{***}	-53.25^{***}	-32.24^{**}	-0.29	-0.15
Electoral democ.	9.28	8.78	-0.22	-0.16	-5.85	-5.00	-0.55	-0.57
FH civil liberties	1.16	0.99	0.02	0.01	-1.35	-1.29		
Parliament							0.03	0.02
Managers							0.02	0.02
Labor force							-0.01	-0.01
Polity2							0.07^{*}	0.07
Ethnic frac.							-0.45	-0.58
Literacy pop.							-1.05	-1.40
Openness							0.93	0.76
Not colony							0.04	0.01
British colony							0.34	0.29
(Intercept)	-59.70^{*}	-41.08	14.57^{***}	13.11^{***}	1908.79^{**}	1762.45^{**}	11.33^{*}	13.60^{**}
R^2	0.79	0.81	0.86	0.87	0.83	0.85	0.66	0.68
$\widehat{MSEP}$	265	259	0.504	0.509	1054	1000	1.069	1.141
N	91	91	97	97	97	97	85	85

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors.

Table 3.2: Weights and coefficients in terms of the variables building the SIGI for female education

	$\hat{\beta}_{PCR}$	w_{PCA}	$\hat{\beta}_{PLSR}$	w_{PLS}
Parental authority 1	−0.62	0.232	−2.19*	0.409
Parental authority 2	−1.98	0.746	−1.78	0.332
Inheritance 1	−1.26	0.475	−2.20*	0.412
Inheritance 2	−1.48	0.560	−3.53**	0.660
Early marriage	−4.35	1.642	−16.63**	3.109
Polygamy 1	0.13	−0.050	−0.95	0.178
Polygamy 2	−2.02	0.762	−4.29**	0.802
Freedom of movement 1	−1.61	0.606	−0.79	0.147
Freedom of movement 2	−3.63	1.368	−3.19	0.596
Freedom of dress 1	−1.35	0.510	0.55	−0.104
Freedom of dress 2	−2.88	1.087	−1.48	0.277
Violence 1	0.92	−0.345	0.77	−0.143
Violence 2	1.11	−0.417	1.81	−0.339
Violence 3	0.44	−0.164	1.90	−0.355
Violence 4	1.22	−0.462	2.69	−0.503
Violence 5	−0.32	0.122	−1.14	0.213
Violence 6	0.88	−0.333	0.70	−0.132
Violence 7	0.81	−0.307	0.58	−0.109
Violence 8	−1.15	0.434	−1.80*	0.337
Violence 9	−1.48	0.558	−2.03	0.379
Female genital mutilation	−2.11	0.794	−6.10**	1.141
Son preference 1	0.07	−0.028	−0.24	0.044
Son preference 2	−1.62	0.611	1.45	−0.271
Son preference 3	−0.85	0.321	−2.46	0.460
Son preference 4	1.92	−0.724	1.01	−0.189
Womens' access to land 1	−1.29	0.486	−2.24*	0.420
Womens' access to land 2	−1.44	0.541	−4.43**	0.829
Womens' access to loan 1	−1.41	0.530	−3.64**	0.680
Womens' access to loan 2	−1.57	0.593	−5.00**	0.934
Womens' access to property other than land 1	−1.44	0.542	−2.23*	0.417
Womens' access to property other than land 2	−1.90	0.715	−3.97**	0.742

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors. For all variables transformed by dummy coding, base category has value 0. Higher value means more gender inequality.

Table 3.3: Weights and coefficients in terms of the variables building the SIGI for child mortality

	$\hat{\beta}_{PCR}$	w_{PCA}	$\hat{\beta}_{PLSR}$	w_{PLS}
Parental authority 1	1.09	0.211	5.75**	0.422
Parental authority 2	3.72	0.723	5.16*	0.379
Inheritance 1	2.46	0.478	5.04*	0.370
Inheritance 2	2.89	0.563	9.27**	0.680
Early marriage	8.75	1.702	40.91***	3.003
Polygamy 1	−0.11	−0.021	2.46	0.181
Polygamy 2	3.77	0.733	9.51**	0.698
Freedom of movement 1	3.13	0.608	1.52	0.112
Freedom of movement 2	6.71	1.304	1.42	0.105
Freedom of dress 1	2.53	0.492	−2.20	−0.161
Freedom of dress 2	5.51	1.072	−0.34	−0.025
Violence 1	−1.93	−0.375	−4.62	−0.339
Violence 2	−1.80	−0.351	−4.24	−0.311
Violence 3	−0.98	−0.190	−5.52*	−0.406
Violence 4	−2.57	−0.500	−5.21*	−0.382
Violence 5	0.53	0.102	−0.11	−0.008
Violence 6	−1.86	−0.362	−2.64	−0.194
Violence 7	−1.75	−0.341	−4.52	−0.332
Violence 8	2.04	0.397	5.17**	0.379
Violence 9	3.03	0.590	10.16	0.746
Female genital mutilation	4.14	0.805	15.66***	1.150
Son preference 1	−0.21	−0.040	1.53	0.112
Son preference 2	3.05	0.592	−5.50*	−0.403
Son preference 3	1.31	0.255	0.75	0.055
Son preference 4	−4.05	−0.788	−6.59	−0.484
Womens' access to land 1	2.67	0.520	5.67**	0.416
Womens' access to land 2	2.65	0.515	11.43**	0.839
Womens' access to loan 1	2.88	0.560	10.13***	0.744
Womens' access to loan 2	2.86	0.557	9.76*	0.716
Womens' access to property other than land 1	2.92	0.567	5.56**	0.408
Womens' access to property other than land 2	2.88	0.561	10.38	0.762

Note: *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Jackknife standard errors. For all variables transformed by dummy coding, base category has value 0. Higher value means more gender inequality.

PCA-based SIGI. On the other hand, both SIGIs based on PCA and PLS are significant in the fertility and corruption regression at 5% or 1% level. More gender inequality increases *fertility* and *corruption*.

Table 3.2 shows the PCR/PLSR coefficients in terms of the variables building the SIGI from the *female education* regression and the weights. No variable has a significant effect in the PCR. On the other hand, high inequality in inheritance, early marriage, high prevalence of polygamy, female genital mutilation, high inequality in women's access to land and properties other than land and high and medium inequality in women's access to loan have significant negative effects on female education in the PLSR. These variables are particularly relevant for the prediction for female education, considering the better prediction performance of the PLSR. A comparison of the PLS weights vis-à-vis the PCA weights show which variables are important to build a composite index relevant to female education. Early marriage and a moderate level of violence against women (Violence 3) are emphasized by the PLS weights, while high level of inequality in parental authority, freedom of movement, freedom of dress, some parts of violence against women (Violence 1, 6 and 7) and stong son preference (Son preference 4) are understated. For medium prevalence of polygamy, medium level of inequality in freedom of dress and low level of son preference (Son preference 1 and 2), the PLS and PCA weights have opposite signs.

Table 3.3 is from the *child mortality* regression. We do not see any significant variables in the PCR, whereas medium inequality in parental authority, high inequality in inheritance, early marriage, high prevalence of polygamy, high level of violence against women (Violence 8), female genital mutilation, medium and high inequality in womens' access to land and medium inequality in women's access to loan and property other than land are significant in the PLSR. These variables can be considered to be important for the prediction for child mortality. The PLS weights emphasize medium level of inequality in parental authority, early marriage and a part of violence against women (Violence 3) and understates freedom of movements and medium high level of son preference (Son preference 3) campared to the PCA weights. For medium prevalence of polygamy, medium and high level of inequality in freedom of dress, a part of violence against women (Violence 5) and low level of son preference (Son preference 1 and 2), the PLS and PCA weights have opposite signs.

For *fertility* and *corruption* regressions, the PLSRs and the PCRs show similar prediction performance, while the PCRs show slightly smaller estimated MSEP. PLSR usually outperforms PCR, because PLS algorithm draws information from outcome variable to enhance prediction. However, too many control variables in fertility and corruption regressions could have caused overfitting. Without the control variables, the PLSR outperforms the PCR for both outcomes. Given the similar performance of the PLSRs and PCRs for these outcome variables, a comparison between the PLSRs and PCRs seems to be not informative. Hence, we do not report the coefficients and weights here, but in Appendix 3.A.

3.5 CPI

In this section we consider the relationship between gender inequality and social institutions and the level of corruption as measured by the CPI. Contuing with our approach, we will use PLS and PCA to assign weights to both the SIGI as well as the CPI. The PLS will consider the relationship between the indicators included in the SIGI and the indicators included in the CPI when deriving the weights. This way we can build a CPI and a SIGI that emphasize particularly the relationship between gender inequality and corruption.

We generate new CPIs for the following reasons. First, Transparency International (2013) uses an average to assign equal weights to the indicators in the CPI. Unless all the indicators are equally informative, such a weighting procedure will deteriorate the quality of the composite index. Therefore, we use PCA and PLS to assign weights, which work either when the largest variations in the variables capture corruption, or when gender inequality is actually related to corruption, which has some variations in the variables in the CPI. Second, many indicators included in the CPI have high proportion of missing values. Too many missing values will introduce unacceptable errors to the composite index and cause failures to imputation. We will drop the variables with high proportion of missing values and work with the remaining. Third, Branisa et al. (2013) take the average of the CPIs from several subsequent years as the outcome variable, which we follow in Section 3.4. The CPIs from subsequent years typically include some same indicators. An average over years will generate a composite index emphasizing the indicators appearing often over years, which are not necessarily informative. For that reasons, each variable is used not more than once as we create the CPI. Fourth, the CPI has two sources, surveys and expert opinions. The CPI puts more weight on surveys than expert opinions by letting survey variables to appear more often in the data matrix compared to expert opinion variables, while it is not clear that the former is more informative. We prepare the data differently, so that surveys and expert opinions are more equally treated. All in all, we use different and arguably improved procedures and also use PLS and PCA to generate weights.

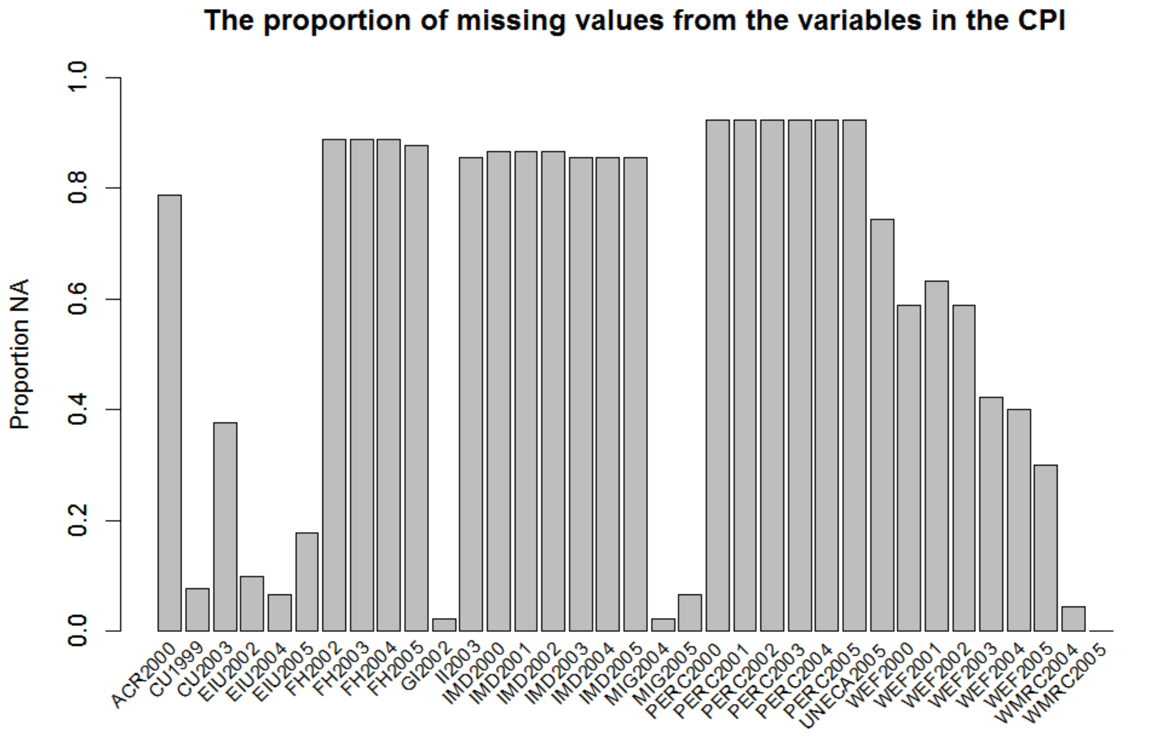
We prepare the data to build the CPI as follows. We work with the variables included in the CPI as scaled by Transparency International (2013). The variables are based on surveys on various types of people with different foci of questions or various expert opinions. The variables are of ordinal nature and transformed to numerical variables. The transformation begins with calculating the ranks of available observations from a variable. The subsample of the CPI from the previous year with the same available observations as the variable are selected, sorted in descending order according to the ranks, and replace the variable. For example, if a variable this year has three observations with a ascending ranking of Germany, France and Italy and the CPI from those countries from the last year are 8, 9.5 and 5, the observations are scaled as 9.5 for Germany, 8 for France and 5 for Italy. The CPI from the previous year takes a value between 0 and 10 with high value meaning less corruption. At the end, the transformed variable again takes a similar scaling as the CPI from the last year. We pool all variables building the CPIs from 2002 to 2005, because we are interested in the level of corruption similar to the time periods of the corruption regression in Section 3.4. Overlapping variables are dropped during the pooling, so that variables appearing more often across years do not get too much emphasis. The CPI from a certain year contains not only variables from the current year, but also lagged variables up to 3 years. The CPI allows lags only for the variables from surveys, but not from the variables from expert opinion. Consequently, the survey variables appear more often than the expert opinion variables in the regressor matrix. When a composite index is built as a linear combination of the columns of the regressor matrix, the survey variables are emphasized simply because they appear more often in the regressor matrix, while it is not clear whether they are more informative than the expert opinion variables. Therefore, when we drop variables during the pooling, we do not distinguish variables from surveys or country experts contrary to the Transparency International (2013). With this procedure, the expert opinion variables are treated more equally important as the survey variables. The pooling approach has a caveat that the variables from different years have slightly different scaling schemes, because the scaling scheme of a year depends on the CPI of the previous year. However, since the distribution of the CPI does not show high volatility for the considered time periods, the pooling will not introduce large changes. At the end we have 90 observations for a regression analysis, which are complete for the variables building the SIGI and control variables. The variables building the CPI have a lot of missing values, which can be seen on the upper part of Figure 3.2. Obviously, imputation is an important issue for this data set.

Transparency International (2013) aggregates the scaled variables to build the CPI, which involves a selection of observations, imputation and weights. Observations which have less than three observed variables are dropped. When there are only small number of indicators available, the quality of the resulting composite index is expected to be low. Then the average over all available columns is taken to build the CPI score. Averaging requires that all indicators are equally important. However, one can expect that the quality of the indicators in the CPI to vary because of the various sources and the different foci of questions. Taking available columns implies an imputation, which requires the assumption that unobserved values are missing at random. This assumption means that the probability for an observation to be missing may depend on observed values, but not missing ones (Schafer, 1999). The CPI data might not satisfy the assumption for the following reasons. Some variables in the CPI data have certain structures in the probability that an observation is missing. For example, the data from Information International cover largely Middle Eastern countries and the data from United Nations Economic Commission for Africa include only African countries. It is questionable whether observed variables contain sufficient information on such structures. Furthermore, the lower part of Figure 3.2 shows the relationship between log GDP and the number of NA of each observation by means of a scatter plot and a fitted line from a simple linear regression. The slope is about -2 and significant at 1% level, which shows that with decreasing GDP, there are more missing values. Considering that many poor countries have high corruption, one can suspect structured missing data pattern.

Transparency International (2013) stretches the distribution of the CPI, so that the variances of the CPI remain similar across different years, which is not relevant for our cross-sectional analysis.

We take the selection of the observations and the imputation method similar to Transparency International (2013), but drop low quality variables and change the weighting procedure to PCA or PLS. We drop variables containing more than 40% of NA, because they can introduce large errors during an imputation. The 15 kept variables are summa-

Figure 3.2: Missing value patterns in the CPI data



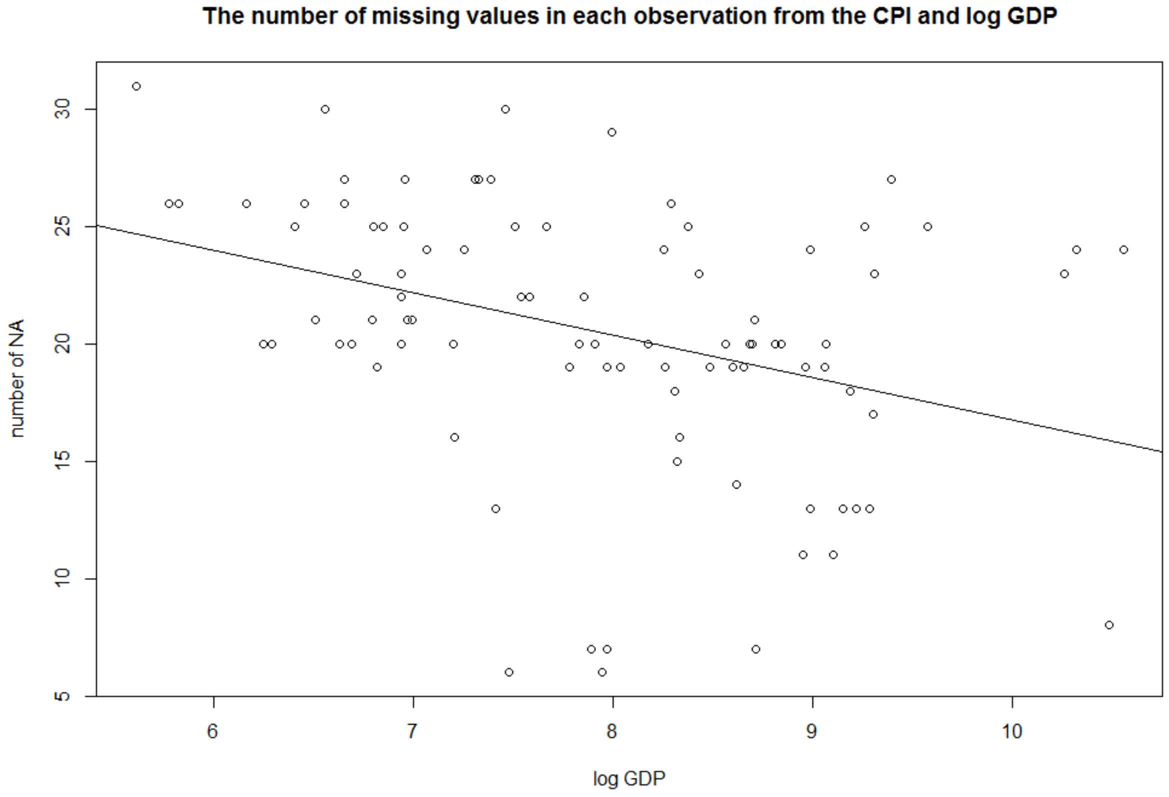


Table 3.4: Linear regressions with the SIGI built by PCA and PLS on the CPI

	CPI	
	$\hat{\gamma}_{PCR}$	$\hat{\gamma}_{PLSR}$
SIGI	-0.92^{**}	-1.06^{*}
log GDP	0.98	1.30^{*}
Parliament	0.09	0.10
Managers	0.11	0.09
Labor force	-0.01	-0.00
Electoral democ.	-0.12	0.61
Polity2	0.13	0.06
SA	-0.77	-0.87
ECA	-6.53^{**}	-4.53
LAC	-5.00^{***}	-2.29
MENA	0.73	1.78
EAP	-3.29	-1.77
Muslim	0.02	-0.29
Christian	0.02	0.31
Ethnic frac.	-1.04	0.24
Literacy pop.	-4.36	-3.76
Openness	5.90^{*}	2.98
Not colony	1.40	1.32
British colony	0.61	1.37
(Intercept)	-6.22	-11.17
R^2	0.44	0.57
$\widehat{MSEP}$	13.460	13.302
N	90	90

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors.

rized in Table 3.7. Then we keep observations which have at least 3 available observations following Transparency International (2013), while no observation is dropped from this procedure. We take the weighted average of all available columns to build the CPI score, where the weights are determined by PCA or PLS (NIPALS, Wold, 1966a; Puwakkatiya-Kankanamage et al., 2014). Under this procedure, the SIGI based on PLS is identified by the maximization of the covariance between the SIGI score and the CPI score built by the NIPALS algorithm. The SIGI based on PCA is not influenced by the NIPALS algorithm, i.e., the solution is same as the usual PCA. Our choice of the NIPALS imputation is motivated by the similarity to the original CPI procedure, one taking a weighted aver-

Table 3.5: Weights and coefficients in terms of variables building the SIGI for the new CPI

	$\hat{\beta}_{PCR}$	w_{PCA}	$\hat{\beta}_{PLSR}$	w_{PLS}
Parental authority 1	-0.21	0.232	0.26	-0.245
Parental authority 2	-0.66^{*}	0.717	-0.65	0.612
Inheritance 1	-0.42	0.458	-0.44	0.418
Inheritance 2	-0.54^{*}	0.588	-0.06	0.056
Early marriage	-1.58^{**}	1.714	-3.15^{*}	2.957
Polygamy 1	-0.00	0.004	0.31	-0.287
Polygamy 2	-0.67^{**}	0.726	-0.85	0.798
Freedom of movement 1	-0.53	0.575	-0.07	0.070
Freedom of movement 2	-1.26	1.362	-0.80	0.752
Freedom of dress 1	-0.46	0.496	-0.10	0.096
Freedom of dress 2	-1.05	1.132	-0.38	0.356
Violence 1	0.34	-0.369	-0.25	0.239
Violence 2	0.45	-0.488	0.27	-0.255
Violence 3	0.17	-0.189	1.26	-1.188
Violence 4	0.43	-0.468	0.66	-0.618
Violence 5	-0.10	0.109	-0.13	0.119
Violence 6	0.33	-0.356	-0.04	0.037
Violence 7	0.31	-0.339	0.13	-0.121
Violence 8	-0.37^{*}	0.404	-0.56	0.525
Violence 9	-0.55	0.601	-0.21	0.200
Female genital mutilation	-0.77^{*}	0.832	-1.17^{*}	1.100
Son preference 1	-0.06	0.067	0.60	-0.564
Son preference 2	-0.56	0.608	0.24	-0.226
Son preference 3	-0.23	0.248	0.24	-0.223
Son preference 4	0.72	-0.782	0.40	-0.380
Womens' access to land 1	-0.46	0.501	-0.29	0.271
Womens' access to land 2	-0.50	0.544	-0.57	0.538
Womens' access to loan 1	-0.52^{*}	0.562	-0.47	0.440
Womens' access to loan 2	-0.54	0.585	-1.31	1.232
Womens' access to property other than land 1	-0.52^{*}	0.564	-0.11	0.100
Womens' access to property other than land 2	-0.54	0.586	-0.82	0.768

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors. For all variables transformed by dummy coding, base category has value 0. Higher value means more gender inequality.

Table 3.6: Weights of the new CPI

	c_{PCA}	c_{PLS}
CU1999	0.617	0.487
CU2003	0.479	0.449
EIU2002	−0.211	0.286
EIU2004	−0.204	0.097
EIU2005	−0.147	−0.238
GI2002	0.008	0.058
MIG2004	−0.325	−0.013
MIG2005	−0.177	−0.188
WEF2000	0.193	0.210
WEF2002	0.141	0.297
WEF2003	0.139	0.277
WEF2004	0.136	0.245
WEF2005	0.133	0.217
WMRC2004	−0.129	−0.155
WMRC2005	−0.123	−0.187

age, another a simple average of the available columns. However, the NIPALS algorithm has the similar weakness that it is not appropriate when the missing data pattern is not random (p18, Nelson, 2002). A deeper investigation on the imputation strategies for the CPI data seems to be fruitable, but we do not pursue it further here.

Table 3.4 shows the model fits using the new CPIs. Both SIGIs have negative effect on the CPIs. The coefficient from the PCR is significant at 5% level, but the coefficient from the PLSR is only marginally significant. It could be that the PLSR has consumed more degrees of freedom (see, Krämer and Sugiyama, 2011), which is followed by an overfitting problem. Nevertheless, even with the different definitions of the CPIs, we find that with more gender inequality, there is more *corruption*. We note that the R^2 and the estimated MSEP from the PLSR and PCR are not comparable, because the outcome variables are constructed differently. The outcome variables are composite indices with different weights.

Table 3.5 shows the coefficients in terms of the variables in the SIGI and the weights used in the *corruption* regression with the new CPIs. Since the prediction performance of

Table 3.7: A summary of the variables building the CPI

	source	name	surveyee	focus of the question
CU1999 CU2003	Columbia University	State Capacity Survey	US-resident country experts (policy analysts, academics and journalists)	Severity of corruption within the state
EIU2002 EIU2004 EIU2005	Economist Intelligence Unit	Country Risk Service and Country Forecast	Expert staff assessment (expatriate)	Assessment of the pervasiveness of corruption (the misuse of public office for private or political party gain) among public officials (politicians and civil servants)
GI2002	Gallup International	Corruption Survey	Senior business people from 15 emerging market economies	How common are bribes to politicians, senior civil servants, and judges and how significant of an obstacle are the costs associated with such payments for doing business?
MIG2004 MIG2005	Merchant International Group	Grey Area Dynamics	Expert staff and network of local correspondents	Corruption, ranging from bribery of government ministers to inducements payable to the “humblest clerk”
WEF2000	World Economic Forum	Global Competitiveness Report	Senior business leaders; domestic and international companies	Undocumented extra payments connected with import and export permits, public utilities and contracts, business licenses, tax payments or loan applications are common/not common.
WEF2002				Questions (in addition to those mentioned above) refer to payments connected to favorable regulations and judicial decisions
WEF2003 WEF2004 WEF2005				Undocumented extra payments or bribes connected with various government functions
WMRC2004 WMRC2005	World Markets Research Centre	Risk Ratings	Expert staff assessment	The likelihood of encountering corrupt officials, ranging from petty bureaucratic corruption to grand political corruption

the PCR and PLSR is not comparable, the PLSR coefficients cannot be considered to be better than PCR coefficients in prediction and a comparison in weights is not informative in building the SIGI relevant to corruption. Therefore, we will focus on the interpretation of each column instead of comparing. Early marriage and high prevalence of polygamy are significant predictors in the PCR and the PCA weights emphasize early marriage, strong restrictions in the freedom of movements and dress. The PLSR shows only marginally significant coefficient estimates and the PLS weights emphasize early marriage, moderate violence (Violence 3), female genital mutilation and high inequality in womens' access to land.

Table 3.6 shows the weights of the CPIs. PCA emphasizes the surveys from Columbia University (CU1999, CU2003) and one expert opinion from Merchant International Group (MIG2004), which shows a counter intuitive negative weight. The surveys from Columbia University are important in PLS as well.

3.6 Conclusions

In this paper, we have built SIGIs using both PLS and PCA to determine the weights and tested whether gender inequality has effects on *female education*, *fertility*, *child mortality* and *corruption*. A model selection is performed to select the treatment of non-metric variables and also non-linear terms of control variables. Our empirical model supports that with more gender inequality, there is higher *fertility* and more *corruption*. On the other hand, for *female education* and *child mortality*, we have have different results depending on whether we use PCA or PLS.

For the *female education* and *child mortality* regressions, PLS brings benefits in terms of prediction compared to PCA. We could see which variables are particularly relevant for the prediction of those outcome variables by comparing the PLSR and PCR coefficients in terms of the variables building the SIGI and weights.

We have created new CPIs with PCA and PLS weights instead of using an average as Transparency International (2013), because it is arguable whether all variables in the CPI are equally important. Additionally, variables are prepared differently to drop variables with large errors and to emphasize each source of variables more equally. We have found a significant effect of the SIGI on the new CPI based on PCA, while for the new CPI based on PLS the effect is only marginally significant. One empirical model supports that with more gender inequality, there is more *corruption*. The NIPALS imputation was employed because it is similar to the imputation procedure of the original CPI, but it is questionable whether the NIPALS algorithm is the best way of imputation for the CPI data. Other imputation approaches can be investigated in the future.

3.A Weights and coefficients from the fertility and CPI regressions

Table 3.8: Weights and coefficients in terms of the variables building the SIGI for fertility

	$\hat{\beta}_{PCR}$	w_{PCA}	$\hat{\beta}_{PLSR}$	w_{PLS}
Parental authority 1	0.04	0.211	0.13**	0.435
Parental authority 2	0.14**	0.723	0.11*	0.386
Inheritance 1	0.09	0.478	0.11*	0.383
Inheritance 2	0.11*	0.563	0.19**	0.658
Early marriage	0.34**	1.702	0.88***	3.027
Polygamy 1	−0.00	−0.021	0.03	0.090
Polygamy 2	0.14**	0.733	0.21**	0.721
Freedom of movement 1	0.12*	0.608	0.03	0.098
Freedom of movement 2	0.26	1.304	0.12	0.400
Freedom of dress 1	0.10	0.492	−0.03	−0.092
Freedom of dress 2	0.21	1.072	0.07	0.239
Violence 1	−0.07	−0.375	−0.01	−0.042
Violence 2	−0.07	−0.351	−0.09	−0.321
Violence 3	−0.04	−0.190	−0.15*	−0.527
Violence 4	−0.10*	−0.500	−0.12*	−0.402
Violence 5	0.02	0.102	0.03	0.096
Violence 6	−0.07	−0.362	−0.08	−0.285
Violence 7	−0.07	−0.341	−0.06	−0.201
Violence 8	0.08**	0.397	0.10**	0.353
Violence 9	0.12	0.590	0.18	0.603
Female genital mutilation	0.16**	0.805	0.35***	1.208
Son preference 1	−0.01	−0.040	−0.04	−0.139
Son preference 2	0.12*	0.592	−0.05	−0.166
Son preference 3	0.05	0.255	−0.01	−0.037
Son preference 4	−0.16	−0.788	−0.19	−0.650
Womens' access to land 1	0.10*	0.520	0.14**	0.482
Womens' access to land 2	0.10	0.515	0.24**	0.829
Womens' access to loan 1	0.11*	0.560	0.20**	0.699
Womens' access to loan 2	0.11	0.557	0.23*	0.792
Womens' access to property other than land 1	0.11*	0.567	0.14**	0.464
Womens' access to property other than land 2	0.11	0.561	0.20*	0.678

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors. For all variables transformed by dummy coding, base category has value 0. Higher value means more gender inequality.

Table 3.9: Weights and coefficients in terms of the variables building the SIGI for the CPI

	$\hat{\beta}_{PCR}$	w_{PCA}	$\hat{\beta}_{PLSR}$	w_{PLS}
Parental authority 1	-0.05	0.233	-0.04	0.130
Parental authority 2	-0.17^{*}	0.730	-0.07	0.222
Inheritance 1	-0.10	0.444	-0.07	0.222
Inheritance 2	-0.14^{*}	0.607	-0.07	0.217
Early marriage	-0.38^{**}	1.622	-1.23^{**}	3.666
Polygamy 1	-0.00	0.018	0.03	-0.093
Polygamy 2	-0.17^{**}	0.736	-0.17	0.500
Freedom of movement 1	-0.15	0.661	-0.06	0.192
Freedom of movement 2	-0.32	1.355	-0.32	0.939
Freedom of dress 1	-0.12	0.533	-0.00	0.002
Freedom of dress 2	-0.27	1.137	-0.19	0.576
Violence 1	0.08	-0.350	-0.22	0.650
Violence 2	0.11	-0.460	0.02	-0.055
Violence 3	0.04	-0.152	0.34	-1.014
Violence 4	0.10	-0.440	0.26	-0.781
Violence 5	-0.01	0.042	0.15	-0.446
Violence 6	0.08	-0.328	-0.05	0.145
Violence 7	0.07	-0.308	-0.16	0.484
Violence 8	-0.09^{*}	0.403	-0.14	0.423
Violence 9	-0.14	0.611	-0.19	0.566
Female genital mutilation	-0.18^{*}	0.780	-0.26^{**}	0.761
Son preference 1	-0.02	0.088	0.11	-0.339
Son preference 2	-0.15	0.630	0.15	-0.449
Son preference 3	-0.12	0.507	-0.23	0.675
Son preference 4	0.17	-0.734	0.09	-0.270
Womens' access to land 1	-0.11	0.492	-0.07	0.203
Womens' access to land 2	-0.13	0.557	-0.26^{*}	0.774
Womens' access to loan 1	-0.13	0.538	-0.20^{*}	0.596
Womens' access to loan 2	-0.14	0.581	-0.35^{*}	1.029
Womens' access to property other than land 1	-0.13	0.555	-0.01	0.044
Womens' access to property other than land 2	-0.14	0.601	-0.34^{**}	0.997

Note: *** p<0.01, ** p<0.05, * p<0.1. Jackknife standard errors. For all variables transformed by dummy coding, base category has value 0. Higher value means more gender inequality.

Bibliography

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Journal of Economic growth*, 8(2):155–194.

Barro, R. and Lee, J.-W. (2013). A new data set of educational attainment in the world, 1950-2010. *Journal of Development Economics*, 104:184–198.

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

Bergh, A. and Karlsson, M. (2010). Government size and growth: Accounting for economic freedom and globalization. *Public Choice*, 142(1-2):195–213.

Booysen, F., Van Der Berg, S., Burger, R., Maltitz, M. V., and Rand, G. D. (2008). Using an asset index to assess trends in poverty in seven sub-saharan african countries. *World Development*, 36(6):1113–1130.

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

Cantaluppi, G. (2012). A partial least squares algorithm handling ordinal variables also in presence of a small number of categories. *arXiv preprint*, arXiv:1212.5049.

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.

Cingranelli, D. L. and Richards, D. L. (2006). The Cingranelli-Richards (CIRI) human rights dataset 2006. url = http://www.humanrightsdata.org/.

Clark, W. C. (2000). Environmental globalization. In Nye, J. S. and Donahue, J. D., editors, *Governance in a globalizing world*, page 86. Brookings Institution Press, Washington, DC.

Correlates of War 2 Project. (2003). Colonial/dependency contiguity data, v3.0. url = http://correlatesofwar.org/.

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

DG Enterprise (2001). Summary Innovation Index. url=http://ec.europa.eu/enterprise/policies/innovation/policy/innovation-scoreboard/.

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

Dreher, A., Gaston, N., and Martens, P. (2008). *Measuring Globalisation: Gauging Its Consequences.* Springer.

Feenstra, R. C., Inklaar, R., and Timme, M. P. (2013). The next generation of the penn world table. available for download at = www.ggdc.net/pwt.

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.

Freedom House (2008). Freedom in the world 2008. url = http://www.freedomhouse.org.

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

Habib, M. and Zurawicki, L. (2002). Corruption and foreign direct investment. *Journal of international business studies*, pages 291–307.

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.

Keohane, R. O. and Nye, J. S. (2000). Introduction. In Nye, J. S. and Donahue, J. D., editors, *Governance in a globalizing world*, pages 1–44. Brookings Institution Press, Washington, DC.

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.

KOF Swiss Economic Institute (2013). KOF Index of Globalization. url = http://globalization.kof.ethz.ch/.

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.

Krämer, N. and Sugiyama, M. (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, 106(494).

Maitra, S. and Yan, J. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying Multivariate Statistical Models*, 79.

Mankiw, N. G., Romer, D., and Weil, D. N. (1992). A contribution to the empirics of economic growth. *The quarterly journal of economics*, 107(2):407–437.

Martens, H. and Martens, M. (2000). Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (plsr). *Food quality and* *preference*, 11(1):5–16.

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.

Monty G. Marshall (2013). Polity IV Project: Political Regime Characteristics and Transitions, 1800-2012. url = http://www.systemicpeace.org/polity/polity4.htm.

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.

Nelson, P. R. C. (2002). *The Treatment Of Missing Measurements In PCA And PLS Models*. PhD thesis, McMaster University.

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.

Norris, P. (2000). Global governance and cosmopolitan citizens. In Nye Jr, J. S. and Donahue, J. D., editors, *Governance in a globalizing world*, pages 173–75. Brookings Institution Press, Washington, DC.

Potrafke, N. (2014). The evidence on globalization. *World Economy*. forthcoming.

Puwakkatiya-Kankanamage, E. H., García-Muñoz, S., and Biegler, L. T. (2014). An optimization-based undeflated pls (oupls) method to handle missing data in the training set. *Journal of Chemometrics*.

Rao, B. B., Tamazian, A., and Vadlamannati, K. C. (2011). Growth effects of a comprehensive measure of globalization with country specific time series data. *Applied Economics*, 43(5):551–568.

Ravallion, M. (2012a). Poor, or just feeling poor? on using subjective data in measuring poverty. on using subjective data in measuring poverty. *World Bank Policy Research Working Paper*, (5968).

Ravallion, M. (2012b). Poverty lines across the world. In Jefferson, P. N., editor, *The Oxford Handbook of the Economics of Poverty*, chapter 3. Oxford University Press.

Rischke, R., Kimenju, S. C., Qaim, M., and Klasen, S. (2014). Supermarkets and the nutrition transition in kenya. *GlobalFood Discussion Papers*, pages ISSN (2192–3248).

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.

Sahn, D. E. and Stifel, D. C. (2000). Poverty comparisons over time and across countries in africa. *World development*, 28(12):2123–2155.

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.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.

Sen, A. (1999). *Development as freedom.* Oxford University Press.

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.

Transparency International (2013). Corruption Perception Index. url = http://www.transparency.org/.

United Nations Development Programme (1995). *Human Development Report.* Oxford University Press, New York.

Wittenberg, M. (2013). The weight of success: The body mass index and economic well-being in southern africa. *Review of Income and Wealth*, 59(S1):S62–S83.

Wold, H. (1966a). Estimation of principal components and related models by iterative least squares. In Krishnaiah, P., editor, *Multiuariate Analysis*, pages 391–420. Academic Press, New York.

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

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. *LECTURE NOTES IN MATHEMATICS*, 973:286–293.

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

World Bank (2008). World development indicators. url = http://data.worldbank.org/data-catalog/world-development-indicators.

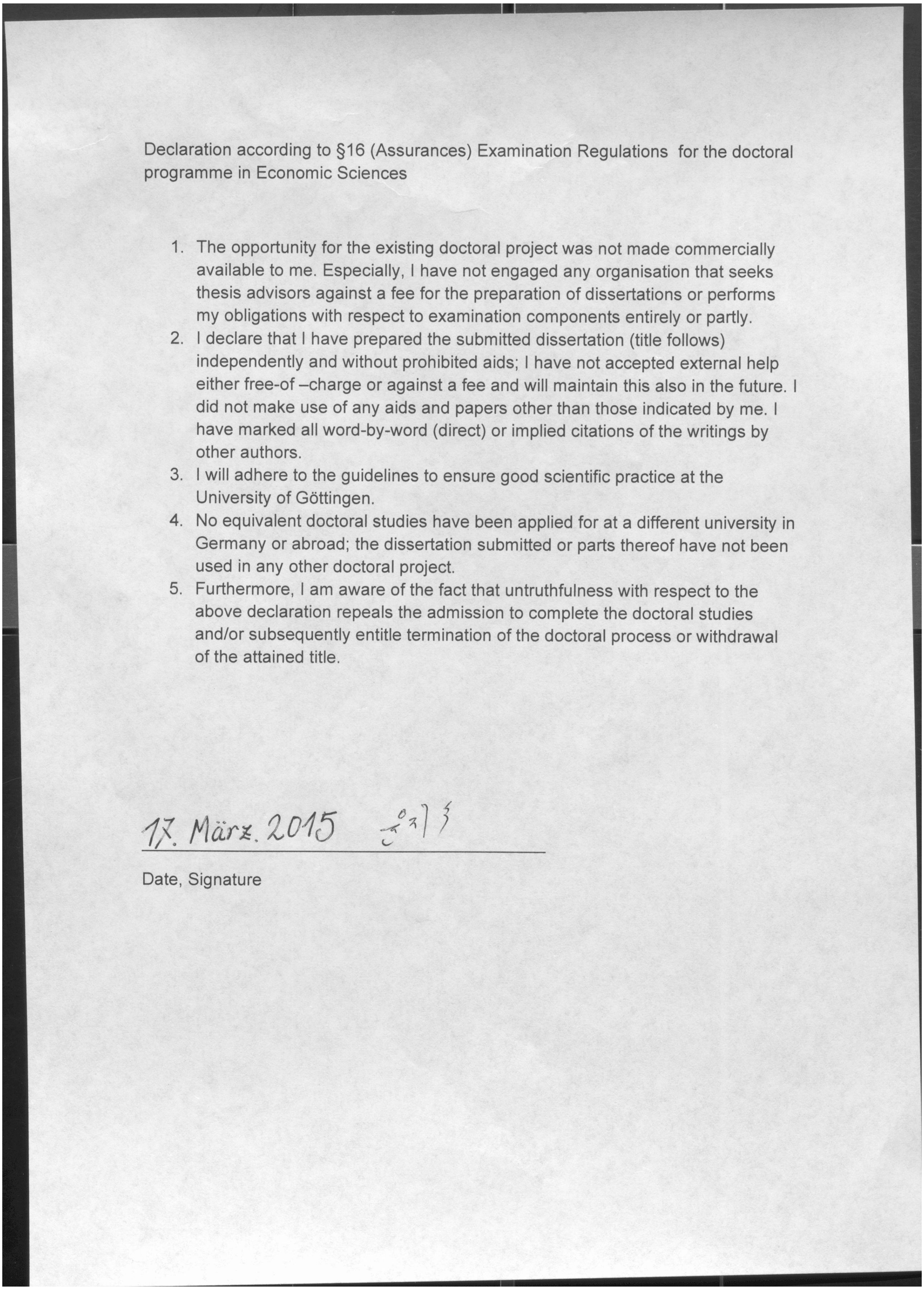
World Bank (2009). GenderStats. url = http://datatopics.worldbank.org/gender/.

World Bank (2013). World Development Indicators. url = http://data.worldbank.org/data-catalog/world-development-indicators.

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

Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432.

Eidesstattliche Erklärung



Declaration according to §16 (Assurances) Examination Regulations for the doctoral programme in Economic Sciences

1. The opportunity for the existing doctoral project was not made commercially available to me. Especially, I have not engaged any organisation that seeks thesis advisors against a fee for the preparation of dissertations or performs my obligations with respect to examination components entirely or partly.
2. I declare that I have prepared the submitted dissertation (title follows) independently and without prohibited aids; I have not accepted external help either free-of –charge or against a fee and will maintain this also in the future. I did not make use of any aids and papers other than those indicated by me. I have marked all word-by-word (direct) or implied citations of the writings by other authors.
3. I will adhere to the guidelines to ensure good scientific practice at the University of Göttingen.
4. No equivalent doctoral studies have been applied for at a different university in Germany or abroad; the dissertation submitted or parts thereof have not been used in any other doctoral project.
5. Furthermore, I am aware of the fact that untruthfulness with respect to the above declaration repeals the admission to complete the doctoral studies and/or subsequently entitle termination of the doctoral process or withdrawal of the attained title.

17. März. 2015

Date, Signature