Courant Research Centre

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

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 171

Composite Indices Based on Partial Least Squares

Jisu Yoon, Stephan Klasen, Axel Dreher,
Tatyana Krivobokova

March 2015

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

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

Composite Indices Based on Partial Least Squares

Jisu Yoon[1, 2] Stephan Klasen[1, 3] Axel Dreher [4, 3, 5, 6, 7, 8]

Tatyana Krivobokova [1, 9]

Georg-August-Universität Göttingen

March 24, 2015

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]Courant Research Center "Poverty, Equity and Growth", Georg-August-Universität Göttingen, Wilhelm-Weber-Str. 2, 37073 Göttingen, Germany

[2]Tel.: +49 (0)551 39 12187 Fax: +49 (0)551-39 14059 E-mail: jisu.yoon@zentr.uni-goettingen.de

[3]Department of Economics, Georg-August-Universität Göttingen, Germany

[4]Heidelberg University, Alfred-Weber-Institute for Economics, Bergheimer Strasse 58, 69115 Heidelberg, Germany

[5]CESifo, Germany

[6]IZA, Germany

[7]KOF Swiss Economic Institute, Switzerland

[8]CEPR, United Kingdom

[9]Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Göttingen, Germany

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, 1966) 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. 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. While not necessarily implying a causal relationship, 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.

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 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 2 we review basic principles of PLS and PCA, various treatments of non-metric variables for these algorithms and conduct a simulation study. Section 3 presents the analysis of the three data sets and the indices we obtain, while we conclude in Section 4.

2 PCA and PLS with Non-metric Variables

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

$$Y = X\beta + \varepsilon, \tag{1}$$

for $Y = (y_1, \ldots, y_n)^t$, $\beta = (\beta_1, \ldots, \beta_k)^t$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$, with $E(\varepsilon) = 0_n$, $\mathrm{cov}(\varepsilon) = \sigma^2 I_n$. 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

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

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 $X^t X$ as weights, while PLS-based indices have weights $X^t Y$.

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.

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)

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

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 $\underline{1}_n = (1, \ldots, 1)^t$ is a n-dimensional vector of ones and P and S as defined in Subsection 2.2. The coefficient vector $\hat{\gamma}_i$ is the OLS coefficient estimates of Y_i on U_i.

Table 1 reports the simulations results.

Table 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

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

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 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.

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.

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).[1] 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 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

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

cycle, we take 4 year averages of all variables.[2] 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

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

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 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

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

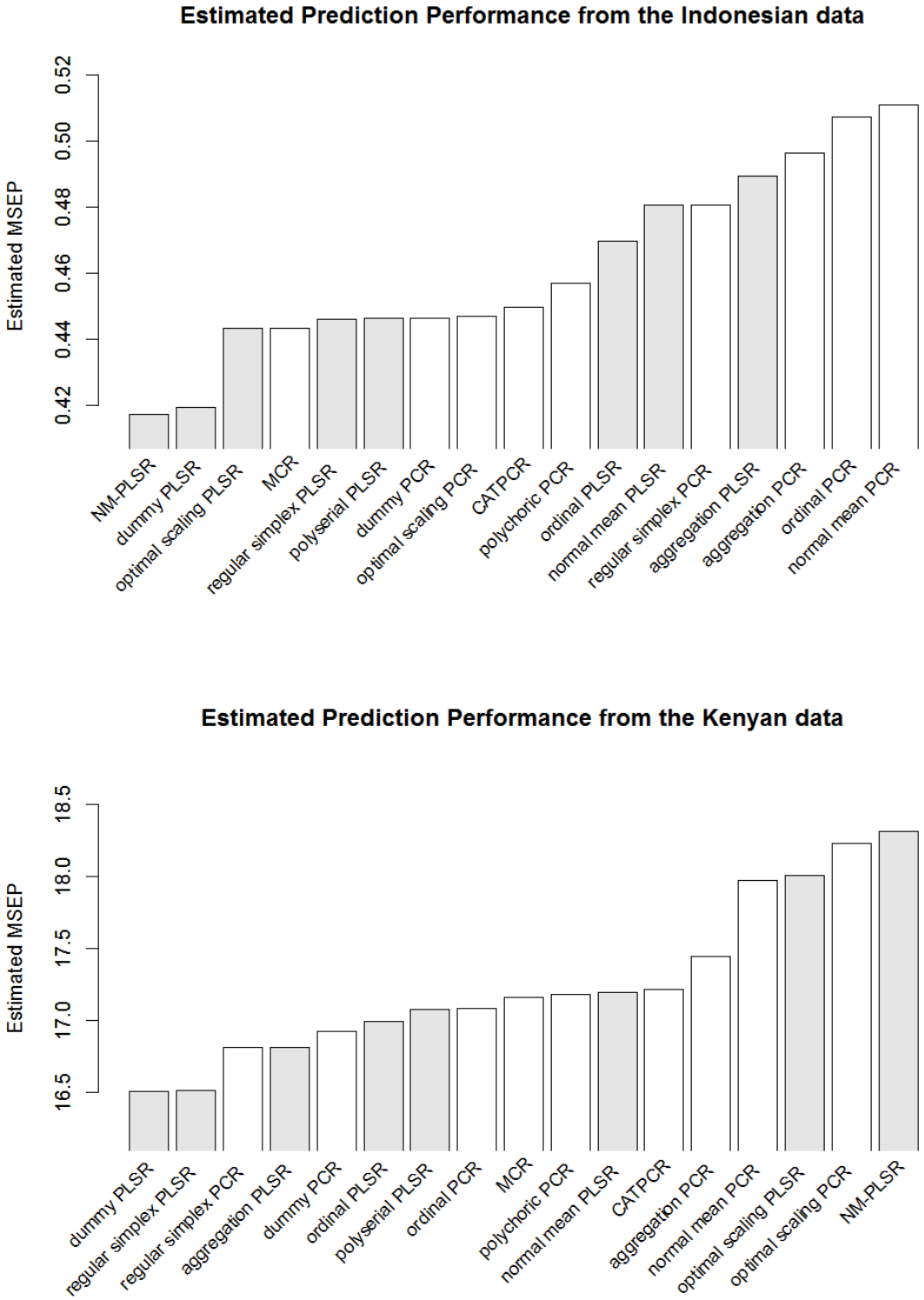
the Indonesian and Kenyan applications. In analogy to the simulation study, PLSR using dummy coding performs excellently. It performs second best for the Indonesian data and best for the Kenyan data.

3.2 Wealth Index with BMI as the Outcome Variable

Table 2 shows the regression coefficients as well as the weights using PCA (left column) and PLS (right columns). 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. 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.

In Table 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

Figure 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 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.

Table 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.

health-related outcome variable improves the predictive performance of the wealth index for other socio-economic outcomes.

3.3 Wealth Index with Expenditure as the Outcome Variable

In Table 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. 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. 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. In analogy to the weights, the PLSR and PCR coefficients show large differences. In the

Table 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.

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 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 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.

3.4 Globalization Index with Growth as the Outcome Variable

Table 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

Table 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) 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.

We turn to our disaggregate analysis in Table 7. As can be seen at the bottom of the table, the R^2 of the PLSR is larger, while the estimated MSEP (using 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.[3] The results differ substantially when we use PLS rather than PCA (right column of Table 7). Almost half of the variables are no longer significant at conventional levels. Regarding actual economic flows, we find that economic growth increases with a country's stock of FDI and portfolio

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

Table 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.

investments (both in percent of GDP on the original scale[4]), 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 increase growth, 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 of treaties signed.

Table 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 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), im-

[4]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.

provements 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 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.

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.

A Descriptions of Variables

Table 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 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 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

References

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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

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

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.

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.

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.

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.

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.

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

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

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