Yoon, Jisu; Klasen, Stephan

Working Paper

An application of partial least squares to the construction of the Social Institutions and Gender index (SIGI) and the Corruption Perception Index (CPI)

Discussion Papers, No. 173

Provided in Cooperation with:
Courant Research Centre 'Poverty, Equity and Growth in Developing and Transition Countries', University of Göttingen

Suggested Citation: Yoon, Jisu; Klasen, Stephan (2015) : An application of partial least squares to the construction of the Social Institutions and Gender index (SIGI) and the Corruption Perception Index (CPI), Discussion Papers, No. 173, Georg-August-Universität Göttingen, Courant Research Centre - Poverty, Equity and Growth (CRC-PEG), Göttingen

This Version is available at:
https://hdl.handle.net/10419/109028

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. 173

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

Jisu Yoon, Stephan Klasen

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

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

Jisu Yoon[1, 2] Stephan Klasen[1, 3]

Georg-August-Universität Göttingen

March 24, 2015

Abstract

In this paper the Social Institutions and Gender Index (SIGI) is constructed with Principal Component Analysis (PCA) and Partial Least Squares (PLS). Using the SIGI, we test the effects of social institutions related to gender inequality on several development outcomes, such as female education, fertility, child mortality and corruption, controlling for relevant determinants. As the measure of corruption we use the Corruption Perception Index (CPI), considering 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.

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

[2]E-mail: jisu.yoon@zentr.uni-goettingen.de

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

1 Introduction

Gender inequlity not only deprives the women of basics 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 2 recapitulates PCA and PLS algorithms. Section 3 discusses the data. Empirical analyses follow in Section 4. In Section 5, we create new CPIs with different weighting schemes. Then we conclude.

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 Data

In this section we explain 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 total 124 observations 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 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.

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.

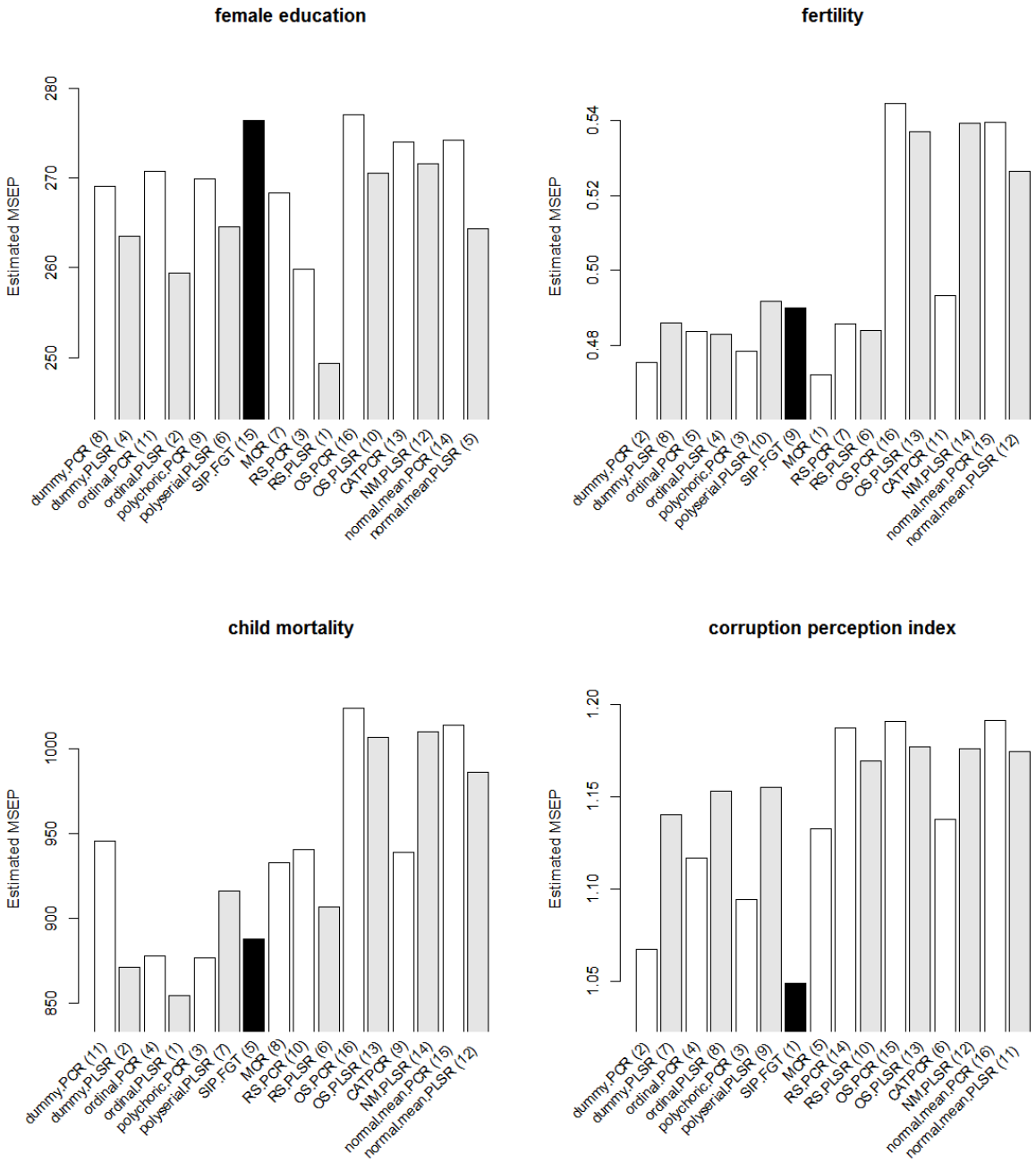
Next, we perform a model selection in terms of various treatments of non-metric variables for PCA and PLS available in the literature. The prediction performance measured by the estimated mean squared error of prediction (MSEP; Mevik and Cederkvist, 2004) via the Jackknife is considered as the model 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. Note that the abbreviation in the parenthesis corresponds to Figure 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. See Chapter **??** for a detailed summary of those methods. 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 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

Figure 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.

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 1 shows the results of the linear regressions for the outcome variables on the SIGI 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 error (Martens and Martens, 2000). The SIGI based on PCA has no significant effect on *female education* and *child mortality*, but the SIGI based on PLS is significant at 5% and 1% level. It suggests that the weights generated by PCA generate a SIGI that has no significant impact on these outcomes, while the SIGI generated by PLS has significant impact, where more gender inequality predicts lower female education and more child mortality. Considering PLS works often better than PCA when the important latent variable has small variations in the indicators, we can suspect large measurement errors are problemetic in the PCA generated 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 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^{*}
$(\text{log GDP})^2$					66.19^{*}	63.40		
$(\text{log GDP})^3$			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 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: 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.

Table 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 PLS weights vis-à-vis PCA weights show which variables are important to build a composite index relevant to female education. Early marriage, medium prevalence of polygamy, medium restriction on freedom of movement and a moderate level of violence against women (Violence 3) are emphasized by PLS, 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 level of inequality in freedom of dress and low level of son preference (Son preference 1 and 2) PLS and PCA weights have opposite signs.

Table 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. PLS weights emphasize medium level of inequality in parental authority and understates medium high level of son preference (Son preference 3). 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 medium level of son preference (Son preference 1), 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. Since a comparison between the PLSRs and PCRs will not show the important variables, which matter for prediction for the models at hand, we do not report the coefficients and weights here, but in Appendix A.

5 CPI

We suggest new aggregation methods to build the CPI in this section. After discussing the motivation of this exercise, we review the data and aggregation method of the original CPI. At the end we generate new CPIs and report the results.

We generate new CPIs for the following reasons. First, the 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 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 weights on surveys than expert opinions, 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.

We prepare the data to build the CPI as follows. We work with the variables included in the CPI as scaled by the 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 decending order according to the ranks, and replace the variable. For example, if a variable this year has three observations with a decending 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 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. 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. However, the variables building the CPI have a lot of missing values, which can be seen on the upper part of Figure 2. Obviously, imputation is an important issue for this data set.

The 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 obviously requires that missing data pattern is random. If a low value leads to a missing value, remaining available variables will show systemetically high values than the actual corruption level. Unfortunately, the data indicates that the missing data pattern is structured. The lower part of Figure 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 indicates that with decreasing

GDP, there are more missing values. Considering that many poor countries have high corruption, one can suspect structured missing data pattern. The 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.

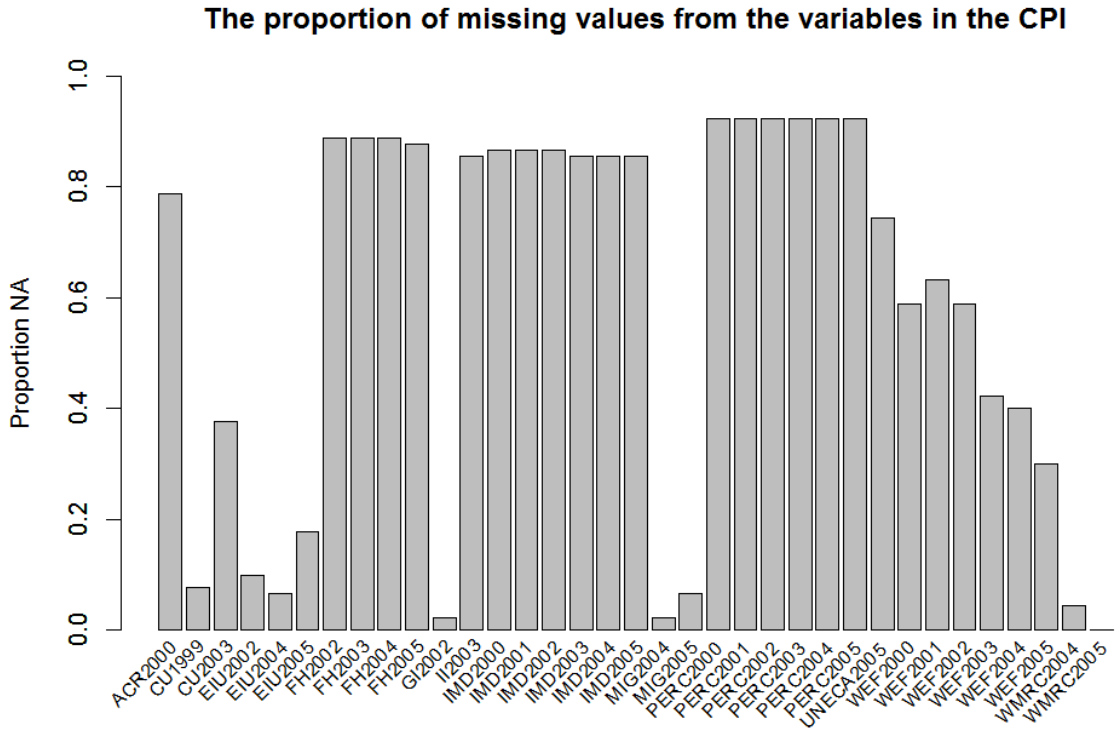
Table 4: Linear regressions with the SIGI built by PCA and PLS on the CPI

	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.

We take the selection of the observations and the imputation method similar to the 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 kept 15 variables are summarized in Table 7. Then we keep observations which have at least 3 available observations

Figure 2: Missing value patterns in the CPI data



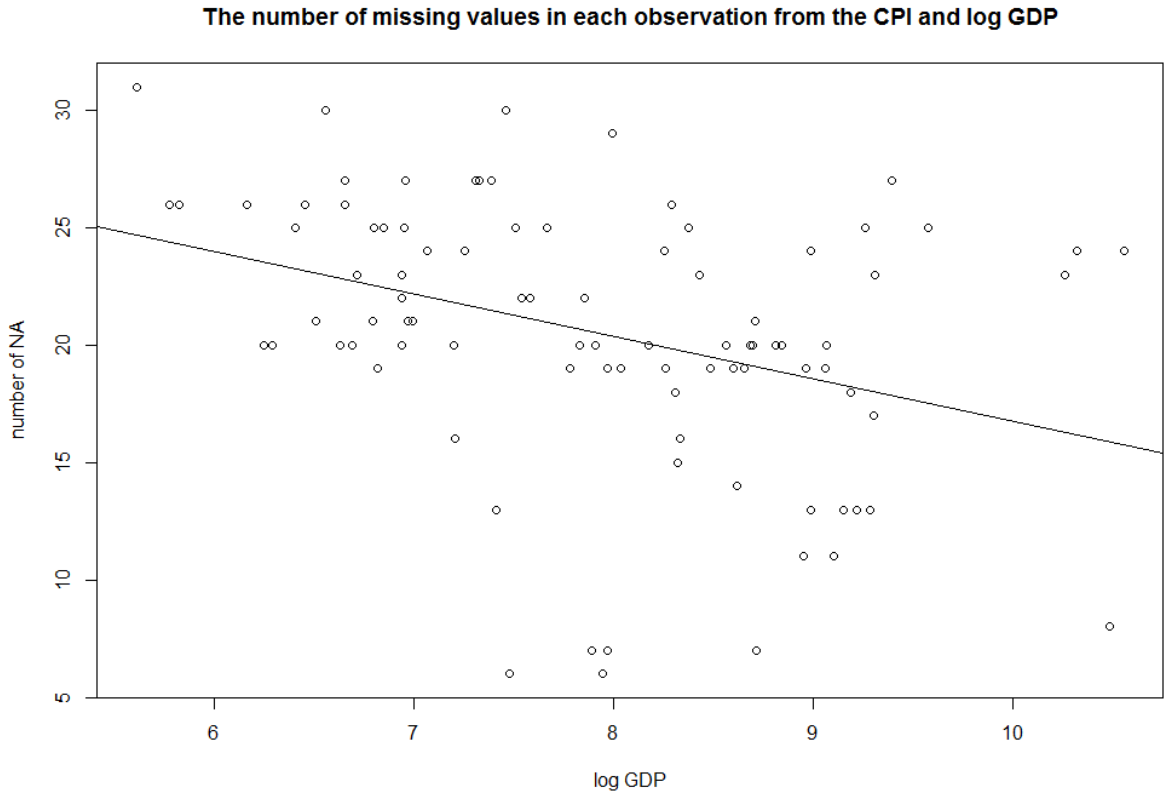


Table 5: Weights and coefficients in terms of variables building the SIGI for the new CPI

	$\hat{\beta}_{PCR}$	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

Table 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

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). Our choice of the NIPALS imputation is motivated by the similarity to the original CPI procedure, one taking a weighted average, 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 non-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 4 shows the model fits using the new CPIs. Both SIGIs have negative effect on the CPIs. The coefficient from the PCR is significant, but the coefficient from the PLSR is only marginally significant. 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.

Table 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

Table 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 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 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.

6 Conclusions

In this paper, we have built the SIGI using both PLS and PCA to determine the weights. Based on the estimated MSEP (via the Jackknife), we have selected dummy coding as the treatment of non-metric variables and also non-linear terms of control variables.

We have tested whether gender inequality has effects on *female education*, *fertility*, *child mortality* and *corruption*. Our empirical model supports that with more gender inequality, there is more *fertility* and *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 *female education* and *child mortality*, PLS brings benefits in terms of prediction compared with PCA. We could see which variables are particularly relevant for the prediction

of those outcome variables by looking at the weights and the coefficients in terms of the variables building the SIGI.

We have created new CPIs with PCA and PLS weights instead of using an average, because it is arguable whether all variables in the CPI are equally important. Variables with too many not available values are dropped to aviod possible large errors. Variables to be included in the new CPIs are selected, so that certain types of variables do not receive too much weights without clear reasons. We have found significant effects of the SIGI on the new CPI based on PCA, while for the new CPI based on PLS the effects are only marginally significant. At least 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 is the best way of imputation for the variables building the CPI. Other imputation approaches to the CPI can be investigated in the future.

A Weights and coefficients from the fertility and CPI regressions

Table 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 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.

References

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

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

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.

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

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

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

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

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

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.

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

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.

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

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.

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*.

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.

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

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.

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.

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/.

Yoon, J. and Krivobokova, T. (2015). Treatments of non-metric variables in partial least squares and principal component analysis. unpublished, tentative publication date.

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.