Attenuation Bias, Measurement Error & Principal Component Analysis

Isaac Liu, Nicolás Martorell & Paul Opheim

May 17, 2021

Abstract

Shorter version of the abstract (I would say 4-5 sentences in a single paragraph max) goes here

Many variables of interest in economics are not directly available as empirical data. Instead, economists often use other variables that are imperfect measurements of the true focus of their analysis. These available variables are known as *proxies* or "variables measured with error", and, if they suffer from classical measurement error, their use causes *attenuation bias* when they are used as independent variables in econometric estimation. Traditionally, instrumental variables are used as a shock of exogeneity to get rid of this bias, but finding truly exogenous variables that satisfy the exclusion restriction is difficult, and so this method can often not be feasibly applied.

As an alternative to dealing with attenuation bias, we propose the use of Principal Component Analysis (PCA) over several variables measured with error. When there are multiple observed variables driven by a single "true" one, we propose to use PCA over these variables to extract the "true" variable. We then use this extracted value and use it in a standard OLS regression, thus providing a solution to attenuation bias that does not require the strong assumptions of instrumental variable analysis.

To show the properties and behaviour of our estimator on large samples under standard assumptions, we present a theoretical framework and a Monte-Carlo analysis. Additionally, we explore a basic empirical application to our method, by estimating the effect of economic development on life expectancy at birth. Since there is no consensus on how to measure economic development, we take a sample of different variables that may measure economic development with error (GDP per capita, GNI per capita, Household Income Per Capita, among others) over which we apply PCA to apply our identification strategy.

Literature

Brief discussion of https://warwick.ac.uk/fac/soc/economics/staff/knagasawa/PartialEffects.pdf, as well as anything else important that comes up on Google Scholar

Theoretical framework

Consider a model where the outcome is denoted by y_i. This outcome depends on a variable of interest denoted by t_i and a vector of covariates denoted by $X_i = (x_{i,1}, x_{i,2}, \dots x_{i,p})'$. Additionally, consider a vector of variables $X_i^* = (x_{i,1}^*, x_{i,2}^*, \dots x_{i,p}^*)'$ that correspond to the covariates X_i but observed with measurement error, where $x_{i,k}^* = x_{i,k} + \eta_{i,k}$ with $\eta_{i,k} \sim iid(0, \sigma_{\eta_k}^2)$, $\mathrm{E}(x_{i,k}'\eta_{i,k}) = 0, \forall i$, $\mathrm{E}(x_{i,k}'\eta_{j,l}) = 0, \forall i \neq j$ and $k \neq l$, and $\mathrm{E}(\eta_{i,k}'\eta_{j,l}) = 0, \forall i \neq j$ and $k \neq l$. Therefore, each $x_{i,k}^*$ suffers from classical measurement error. Note that $\mathrm{E}(x_{i,k}) = \mathrm{E}(x_{i,k}^*) = \mu_{x_k}$ and that $\mathrm{V}(x_{i,k}) = \sigma_{x_k}^2$ while $\mathrm{V}(x_{i,k}^*) = \sigma_{x_k}^2 + \sigma_{\eta_k}^2 \geq \sigma_{x_k}^2$.

Data Generating Process

Assume that the outcome y_i is determined by the following Data Generation Process (DGP):

$$y_i = \gamma t_i + X_i'\beta + \epsilon_i \tag{1}$$

where γ is the parameter of the variable of interest t_i, $\beta = (\beta_1, \beta_2, \dots \beta_p)'$ is the vector of the parameters of the covariates X_i including a constant and $\epsilon_i \sim \mathrm{iid}(0, \sigma_\epsilon^2)$. Under this specification, the coefficients are such that:

$$\begin{pmatrix} \gamma \\ \beta \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{yX} \end{pmatrix} \tag{2}$$

Suppose that the econometrician has access to t_i but, instead of X_i she observes X_i^*. Then, she specifies the following linear model

$$y_i = \gamma^* t_i + X_i^{*\prime}\beta^* + \zeta_i \tag{3}$$

the coefficients would be such that

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} \sigma_t^2 & \Sigma_{tX^*} \\ \Sigma_{X^*t} & \Sigma_{X^*} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{yX^*} \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X + \Sigma_\eta \end{pmatrix}^{-1} \begin{pmatrix} \sigma_t^2 & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{5}$$

To the see the implications of the of this measurement error in the covariates, consider a simple case where the DGP depends only of the variable of interest and a covariate such that:

$$\begin{pmatrix} \gamma \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{6}$$

and with $\sigma_t^2 = \Sigma_X = \Sigma_\eta = 1$ while $\Sigma_{Xt} = 0.6$. Then

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} \gamma^* \\ \beta^* \end{pmatrix} = \begin{pmatrix} 1.37 \\ 0.39 \end{pmatrix}$$

Clearly, both coefficients shows bias when the econometrician assumes a DGP with X_i^*: while there is attenuation bias on the coefficient of the covariate, the coefficient of the variable of interest is biased upward given that some of the effect of the covariates is “omitted” given this attenuation.

Instrumental Variables Regression as a Bias-Correction Method

The classical solution for the measurement-error induced bias in econometrics has been the usage of instrumental variables. Suppose an instrument Z_i that satisfies the relevance condition $\mathrm{E}(Z_i'X_i) \neq 0$ and $\mathrm{E}(Z_i't_i) \neq 0$, and also the exclusion restriction $\mathrm{E}(Z_i'\epsilon_i) = \mathrm{E}(Z_i'\zeta_i) = \mathrm{E}(Z_i'\eta_{i,k}) = 0$, for all i and k. Then premultiplying by Z_i we have

$$Z_i'y_i = Z_i'\gamma^* t_i + Z_i'X_i^{*\prime}\beta^* + Z_i'\zeta_i \tag{7}$$

and so

$$\begin{pmatrix} \gamma^{IV} \\ \beta^{IV} \end{pmatrix} = \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} + \Sigma_{Z\eta} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{8}$$

$$= \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{Zt} & \Sigma_{ZX,Zt} \\ \Sigma_{Zt,ZX} & \Sigma_{ZX} \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} \gamma^{IV} \\ \beta^{IV} \end{pmatrix} = \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{10}$$

However, finding a reliable source of exogeneity is difficult, and it is impossible to conclusively prove a suitable exclusion restriction. The use of IV as a bias-correction method is thus often unfeasible.

Principal Component Regression as Bias-Correction Method

Alternatively, we propose an alternative bias-correction method for when there are several mismeasured variables for each covariate; that is, when we have more than one $x^*_{i,k}$ for every $x_{i,k}$. Given that in all the mismeasured variables the underlying value is the real value, one could think of extracting the underlying true $x_{i,k}$ through a linear combination of the different $x^*_{i,k}$. Then, we could treat all the $x^*_{i,k}$ as variables that share components as follows:

$$h_j = \underset{h'h=1, h'h_1=0, \ldots, h'h_{j-1}=0}{\text{argmax}} \text{var}\left[h'X^*_k\right] \tag{11}$$

where h_j is the eigenvector of Σ associated with the j^{th} ordered eigenvalue λ_j of $\Sigma_{X^*_k}$, and the principal components of X^*_k are $U_j = h'_j X^*_k$, where h_j is the eigenvector of Σ associated with the j^{th} ordered eigenvalue λ_j of Σ.

Under our assumptions, the vector of mismeasured values X^*_k of $x_{i,k}$, share only one principal component which is precisely $x_{i,k}$. Then, we only have one principal component, $x_{i,k}$, and so the $x_{i,k}$ is such that

$$x_{i,k} = h'_k X^*_k \tag{12}$$

Finally, we could then retrieve the vector of true variables X_i

$$X_i = HX^*_i \tag{13}$$

where H is a matrix such that

$$H = \begin{pmatrix} h_1 & 0 & 0 & \ldots & 0 \\ 0 & h_2 & 0 & \ldots & 0 \\ \vdots & \ddots & h_3 & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots \ddots & h_p \end{pmatrix}$$

and h_k is the vector of eigenvalues for the variable $x_{i,k}$.

Our new linear model then becomes

$$y_i = \gamma^{PCR} t_i + HX^{*\prime}_i \beta^{PCR} + \epsilon_i \tag{14}$$

where the coefficients are as follows

$$\begin{pmatrix} \gamma^{PCR} \\ \beta^{PCR} \end{pmatrix} = \begin{pmatrix} \sigma^2_t & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{yt} \\ \Sigma_{y,HX^*} \end{pmatrix} \tag{15}$$

$$= \begin{pmatrix} \sigma^2_t & \Sigma_{t,HX^*} \\ \Sigma_{HX^*,t} & \Sigma_{HX^*} \end{pmatrix}^{-1} \begin{pmatrix} \sigma^2_t & \Sigma_{tX} \\ \Sigma_{Xt} & \Sigma_X \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{16}$$

$$= \begin{pmatrix} \gamma \\ \beta \end{pmatrix} \tag{17}$$

where the last equality comes from (13).

Properties of the Estimator: Monte Carlo Simulations

We then complement our theoretical analysis by using Monte Carlo Simulation to analyze the effects of using Principal Components Regression as a method of bias correction. For these simulations, we assume that the true DGP for the data is:

$$y_i = \beta_1 x_i + \beta_2 z_i + u_i$$

... where x_i and z_i are single variables drawn from $\mathcal{N}(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$, where ρ is some covariance between our main variable of interest (x_i) and the covariate (z_i). The u_i is drawn from a white noise distribution($\mathcal{N}(0,1)$) that is uncorrelated with both x_i and z_i. We then assume (as with the theoretical analysis) that z_i is not directly observable and instead the researchers only have access to p many measurements $z^*_{i,j}$ where $z^*_{i,j} = z_i + \eta_j$ where η_j is drawn from a white noise distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ where $\mathbf{0}$ is a p-vector and Σ is a diagonal p by p matrix with only 1s on the diagonal.

In our simulations, we assume default values of $\rho = 0.5$, $\beta_1 = \beta_2 = 1$, and $p = 5$. We then vary each factor while holding the others fixed, and perform 1,000 simulations of the DGP followed by an OLS regression on either the PCA value from the p measurements of the true z_i, or on a single one of the measurements of z_i. For each simulation, we generate 100 observations of y_i, x_i,etc. The results for each range of parameters can be seen in Appendix 1.

We can see from these simulations that using PCA on several covariates in order to create an estimate of the single latent covariate driving each measurement noticeably outperforms using a single one of those measurements as a variable, with one notable exception. When the covariance between x_i and z_i is equal to $0, -1$, or 1 then there is no notable improvement from using the PCA-extracted latent variable. Both the average coefficient on β_1 obtained when including the PCA output in the regression, and the mean absolute percentage error obtained on the 1,000 simulations are both much closer to the target values with the PCA-based regression than with the single measurement regression.

However, the performance advantages that we see from using PCA could be driven by the benefit of having multiple measurements of our true covariate of interest, as opposed to any special advantages from PCA specifically. We test this question by comparing the estimated β^*_1 in our PCA regressions with the estimated β^*_1 when we include all p measurements as separate covariates in the regression, and the β^*_1 obtained when the covariate is the mean of all p measurements of the true covariate. The results from these regressions can be seen in Appendix 2.

As one can see from these results, there does not seem to be a noticeable difference between these three regression methods (across any values of β_1, β_2, p, and ρ. Thus, our simulations suggest that there are major benefits to having multiple measurements of a latent covariate of interest, but that using PCA, taking the average of these measurements, and including all measurements as separate covariates seem to give similar benefits to the performance of the regression.

Application: Government Share of Healthcare Spending and Life Expectancy

Explain economic importance/interesting-ness of the chosen application

Explain how GDP/economic development is measured with error

It is very difficult to find an instrumental variable for economic development which satisfies a reasonable exclusion restriction.

In the left column in the table below I first regress the life expectancy at birth for all individuals in a given country and year on a measure of government spending as a share of total health expenditure. In the middle column I include the economic controls/covariates of GDP per capita (PPP), GNI per capita (PPP), Survey Mean Income/Consumption Per Capita, ILO GDP per person employed, and Net Foreign Assets Per Capita, all from the World Bank. In the rightmost column I instead use the first principal component combining these covariates.

I standardize all variables by subtracting the mean and dividing by the standard deviation, linearly interpolate data between known observations, and remove country-years with missing values for any of the economic indicators.

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
Govt. Share of Health Exp.	0.613***	0.308***	-0.016	0.343***	-0.012
	(0.019)	(0.020)	(0.019)	(0.019)	(0.018)
Covariates	None	Econ Indicators	Econ Indicators	PCs	PCs
Fixed Effects	No	No	Yes	No	Yes
Observations	1,799	1,799	1,799	1,799	1,799
R^2	0.376	0.586	0.987	0.536	0.987
Adjusted R^2	0.375	0.582	0.985	0.535	0.985
Residual Std. Error	0.791	0.646	0.122	0.682	0.123
F Statistic	1081.530***	168.157***	922.656***	1036.417***	1114.735***

Note: *p<0.1; **p<0.05; ***p<0.01
All variables are standardized

With Ginis instead of econ controls

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
Govt. Share of Health Exp.	0.697***	0.652***	0.028	0.692***	0.026
	(0.034)	(0.040)	(0.024)	(0.037)	(0.023)
Covariates	None	Ginis	Ginis	PCs	PCs
Fixed Effects	No	No	Yes	No	Yes
Observations	322	322	322	322	322
R^2	0.566	0.596	0.999	0.566	0.999
Adjusted R^2	0.565	0.583	0.998	0.564	0.998
Residual Std. Error	0.596	0.583	0.040	0.597	0.040
F Statistic	417.382***	45.905***	1421.063***	208.238***	5197.564***

Note: *p<0.1; **p<0.05; ***p<0.01
All variables are standardized.

Health share of gdp

	Life Expectancy at Birth (Years)				
	(1)	(2)	(3)	(4)	(5)
GDP Share of Health Exp.	0.258***	0.184***	-0.024	0.144***	-0.020
	(0.023)	(0.019)	(0.031)	(0.017)	(0.030)
Covariates	None	Econ Indicators	Econ Indicators	PCs	PCs
Fixed Effects	No	No	Yes	No	Yes
Observations	1,799	1,799	1,799	1,799	1,799
R^2	0.067	0.553	0.987	0.467	0.987
Adjusted R^2	0.066	0.549	0.985	0.467	0.985
Residual Std. Error	0.967	0.672	0.122	0.730	0.123
F Statistic	128.396***	146.946***	772.041***	787.819***	384.737***

Note: *p<0.1; **p<0.05; ***p<0.01
All variables are standardized.

Conclusion

			ρ Value		
	-1	-0.5	0	0.5	1
			Coefficient on Main Variable		
PCA	-0.006	0.900	0.996	1.105	2.009
	(0.238)	(0.120)	(0.111)	(0.121)	(0.242)
Single Measurement	-0.002	0.720	0.998	1.280	2.003
	(0.142)	(0.130)	(0.127)	(0.129)	(0.147)
			Absolute Percentage Error		
PCA	100.6%	12.7%	8.9%	13.1%	100.9%
	(23.8 ppts)	(9.1 ppts)	(6.6 ppts)	(9.3 ppts)	(24.2 ppts)
Single Measurement	100.2%	28.1%	10.2%	28.2%	100.3%
	(14.2 ppts)	(12.7 ppts)	(7.6 ppts)	(12.6 ppts)	(14.7 ppts)
Observations	1,000	1,000	1,000	1,000	

		True β_1		
	0.1	1	10	100
		Coefficient on Main Variable		
PCA	0.207	1.105	10.104	100.117
	(0.121)	(0.121)	(0.123)	(0.124)
Single Measurement	0.383	1.280	10.278	100.289
	(0.128)	(0.129)	(0.131)	(0.133)
		Absolute Percentage Error		
PCA	131.1%	13.1%	1.3%	0.1%
	(95.1 ppts)	(9.3 ppts)	(0.9 ppts)	(0.1 ppts)
Single Measurement	283.6%	28.2%	2.8%	0.3%
	(126.6 ppts)	(12.6 ppts)	(1.3 ppts)	(0.1 ppts)
Observations	1,000	1,000	1,000	1,000

	True β_2			
	0.1	1	10	100
	Coefficient on Main Variable			
PCA	1.018	1.105	2.112	12.171
	(0.115)	(0.121)	(0.477)	(4.555)
Single Measurement	1.034	1.280	3.865	29.751
	(0.107)	(0.129)	(0.703)	(7.231)
	Absolute Percentage Error			
PCA	9.4%	13.1%	111.6%	1,119.6%
	(7.0 ppts)	(9.3 ppts)	(47.0 ppts)	(449.4 ppts)
Single Measurement	8.9%	28.2%	286.5%	2,875.1%
	(6.8 ppts)	(12.6 ppts)	(70.3 ppts)	(723.1 ppts)
Observations	1,000	1,000	1,000	1,000

	Number of p			
	5	10	20	50
	Coefficient on Main Variable			
PCA	1.105	1.066	1.033	1.022
	(0.121)	(0.122)	(0.119)	(0.117)
Single Measurement	1.280	1.283	1.282	1.292
	(0.124)	(0.129)	(0.131)	(0.167)
	Absolute Percentage Error			
PCA	13.1%	11.1%	10.0%	9.4%
	(9.3 ppts)	(8.3 ppts)	(7.3 ppts)	(7.3 ppts)
Single Measurement	28.2%	28.5%	28.3%	29.3%
	(12.6 ppts)	(12.6 ppts)	(12.6 ppts)	(12.7 ppts)
Observations	1,000	1,000	1,000	1,000

	ρ *Value*				
	-1	-0.5	0	0.5	1
			Coefficient on Main Variable		
PCA	-0.006	0.900	0.996	1.105	2.009
	(0.238)	(0.120)	(0.111)	(0.121)	(0.242)
All Measurements	-0.007	0.904	0.996	1.100	2.011
	(0.249)	(0.122)	(0.112)	(0.124)	(0.249)
Average of Measurements	-0.005	0.905	0.996	1.100	2.010
	(0.243)	(0.120)	(0.110)	(0.121)	(0.246)
			Absolute Percentage Error		
PCA	100.6%	12.7%	8.9%	13.1%	100.9%
	(23.8 ppts)	(9.1 ppts)	(6.6 ppts)	(9.3 ppts)	(24.2 ppts)
All Measurements	100.7%	12.6%	9.0%	12.9%	101.1%
	(24.9 ppts)	(9.0 ppts)	(6.7 ppts)	(9.3 ppts)	(24.9 ppts)
Average of Measurements	100.5%	12.4%	8.9%	12.8%	101.0%
	(24.3 ppts)	(8.9 ppts)	(6.6 ppts)	(9.2 ppts)	(24.6 ppts)
Observations	1,000	1,000	1,000	1,000	

	True β_1			
	0.1	1	10	100
		Coefficient on Main Variable		
PCA	0.207	1.105	10.104	100.117
	(0.121)	(0.121)	(0.123)	(0.124)
All Measurements	0.201	1.100	10.098	100.11
	(0.123)	(0.124)	(0.126)	(0.127)
Average of Measurements	0.202	1.100	10.098	100.111
	(0.121)	(0.121)	(0.123)	(0.124)
		Absolute Percentage Error		
PCA	131.1%	13.1%	1.3%	0.1%
	(95.1 ppts)	(9.3 ppts)	(0.9 ppts)	(0.1 ppts)
All Measurements	128.9%	12.9%	1.3%	0.1%
	(93.6 ppts)	(9.3 ppts)	(1.0 ppts)	(0.1 ppts)
Average of Measurements	127.9%	12.8%	1.3%	0.1%
	(93.1 ppts)	(9.2 ppts)	(0.9 ppts)	(0.1 ppts)
Observations	1,000	1,000	1,000	1,000

	True β_2			
	0.1	1	10	100
	Coefficient on Main Variable			
PCA	1.018	1.105	2.112	12.171
	(0.115)	(0.121)	(0.477)	(4.555)
All Measurements	1.02	1.100	2.067	11.664
	(0.118)	(0.124)	(0.477)	(4.519)
Average of Measurements	1.017	1.100	2.061	11.625
	(0.115)	(0.121)	(0.470)	(4.415)
	Absolute Percentage Error			
PCA	9.4%	13.1%	111.6%	1,119.6%
	(7.0 ppts)	(9.3 ppts)	(47.0 ppts)	(449.4 ppts)
All Measurements	9.7%	12.9%	107.1%	1,069.3%
	(7.0 ppts)	(9.3 ppts)	(46.8 ppts)	(445.0 ppts)
Average of Measurements	9.4%	12.8%	106.5%	1,065.2%
	(6.9 ppts)	(9.2 ppts)	(46.0 ppts)	(435.0 ppts)
Observations	1,000	1,000	1,000	1,000

	Number of p			
	5	10	20	50
	Coefficient on Main Variable			
PCA	1.105	1.066	1.033	1.022
	(0.121)	(0.122)	(0.119)	(0.117)
All Measurements	1.100	1.061	1.025	1.010
	(0.124)	(0.129)	(0.131)	(0.167)
Average of Measurements	1.100	1.060	1.026	1.015
	(0.121)	(0.122)	(0.119)	(0.117)
	Absolute Percentage Error			
PCA	13.1%	11.1%	10.0%	9.4%
	(9.3 ppts)	(8.3 ppts)	(7.3 ppts)	(7.3 ppts)
All Measurements	12.9%	11.4%	10.7%	13.2%
	(9.3 ppts)	(8.5 ppts)	(7.9 ppts)	(10.2 ppts)
Average of Measurements	12.8%	10.9%	9.8%	9.3%
	(9.2 ppts)	(8.2 ppts)	(7.2 ppts)	(7.2 ppts)
Observations	1,000	1,000	1,000	1,000

Figure 1: Correlations Between Covariates and Life Expectancy

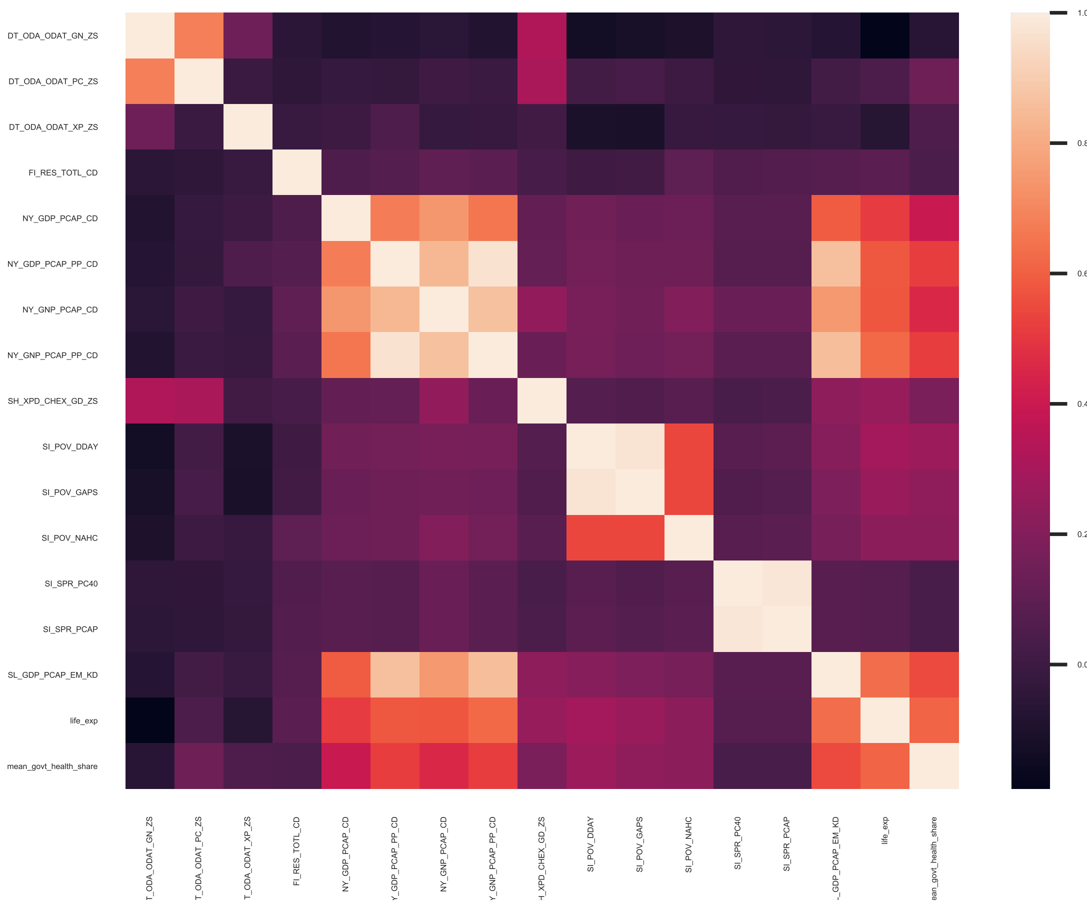

Figure 2: Economic Measures PCA Loadings

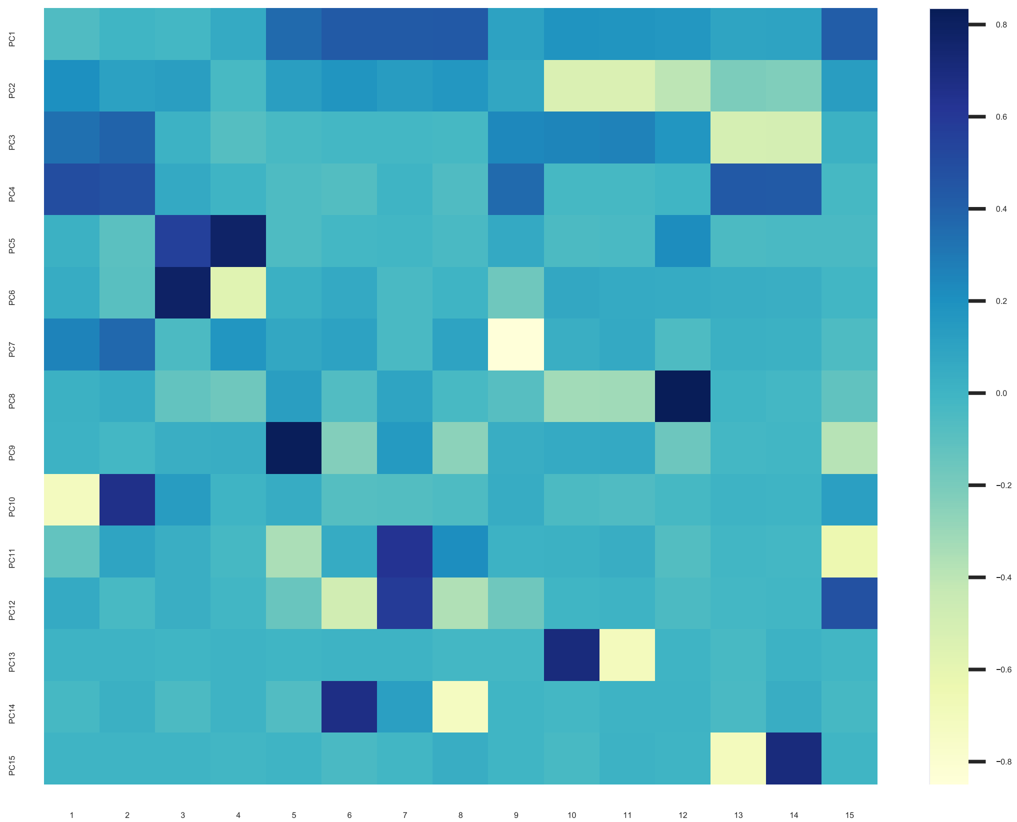

Figure 3: Economic Measures PCA Share of Variance Explained