Spurious Principal Components

Philip Hans Franses
Eva Janssens

EI2017-31
Econometric Institute
Erasmus School of Economics

Abstract

The Principal Component Regression is often used to forecast macroeconomic variables when there are many predictors. In this letter, we argue that it makes sense to pre-whiten the predictors before including these in a PCR. With simulation experiments, we show that without such pre-whitening, spurious principal components can appear, and that these can become spuriously significant in a PCR. With an illustration to annual inflation rates for five African countries, we show that non-spurious principal components can be genuinely relevant in empirical forecasting models.

Key words: Principal Component Regression; Pre-whitening; Spurious Regressions
JEL codes: C52

This version: November 2017

Correspondence: Econometric Institute, Erasmus School of Economics, PO Box 1738, NL-3000 DR, the Netherlands, franses@ese.eur.nl

Introduction and motivation

The Principal Component Regression (PCR) is a frequently considered model to forecast macroeconomic variables when there are many predictors, see Stock and Watson (1999, 2002), Bernanke, Boivin and Eliasz (2005), Heij, van Dijk, and Groenen (2011) and many others. The idea of the PCR is that the predictors are summarized in a few Principal Components, and that these new variables enter as explanatory variables in a regression model. When summarizing the predictors, it is typical practice to consider growth rates of the predictors in case of unit roots, but otherwise the variables are usually included as they are. In this letter, we recommend to pre-whiten all predictors, that is, to fit for example autoregressive models to the data, and use the residuals as the new predictors in Principal Components Analysis (PCA). When the PCA results for raw and pre-whitened data are similar, one may well have found non-spurious Principal Components.

We base our recommendation on a few simulation experiments, which show that without such pre-whitening one runs the risk of finding spurious Principal Components, and finding spuriously significant newly created regressors in the PCR. The arguments why one can obtain spurious effects are the same as those echoed in Yule (1926), Ames and Reiser (1961) and, of course, Granger and Newbold (1974).

An illustration of how a PCR can look like in case of spurious and non-spurious Principal Components is also given.

Simulation experiments

Consider the creation of four time series variables, using the Data Generating Process (DGP):

$$w_t = \alpha_w w_{t-1} + \varepsilon_t^w, \quad \varepsilon_t^w \sim N(0,1)$$

$$x_t = \alpha_x x_{t-1} + \varepsilon_t^x, \quad \varepsilon_t^x \sim N(0,1)$$

$$y_t = \alpha_y y_{t-1} + \varepsilon_t^y, \quad \varepsilon_t^y \sim N(0,1)$$

$$z_t = \alpha_z z_{t-1} + \varepsilon_t^z, \quad \varepsilon_t^z \sim N(0,1)$$

Hence, there are four independent variables, each generated as a first order autoregression. The error terms are all independent draws from a standard normal distribution. The starting values are always equal to 0. In the simulations, *t* will run from 1 to 50, or 100, or 500.

First, we create Principal Components for the variables $x_t, y_t, and\ z_t$, which is done based on the correlation matrix of these three variables. This implies that the sum of the eigenvalues is equal to 3. If the three variables each would be a white noise process, then the estimated eigenvalues should all be about equal to 1. However, when the autoregressive parameter deviates further away from 0 and approaches 1, we may expect that there will appear spurious non-zero correlations across the variables, as already demonstrated in Yule (1926), and hence we may expect that the first eigenvalue will deviate away from 1.

A confirmation of these expectations is summarized in Table 1. The cells in the first panel present the average value of the first eigenvalue and the standard deviation, across 10000 replications. It is clear that the larger the autoregressive parameter gets, the larger is the first eigenvalue. When the sample size increases, the deviation away from 1 gets smaller, but not that much. In the second panel, we report the frequency of 5% significant parameters, associated with the first principal component in the PCR. There, we additionally have that

$$w_t = \alpha_w w_{t-1} + \varepsilon_t^w, \quad \varepsilon_t^w \sim N(0,1),$$

with $\alpha_w = \alpha$ like the other three variables, and where the PCR is

$$w_t = \mu + \rho w_{t-1} + \beta pc_{t-1} + \varepsilon_t,$$

with pc_{t-1} denoting the first lag of the first principal component. Clearly, there are more than 5% significant β parameters, but the spurious effects tend to disappear as we let the sample size increase.

Table 2 presents similar information as Table 1, although now all variables have been pre-whitened, that is, for all variables we first estimate a first order autoregression, and then we proceed with the residuals. Hence, we now first run the regressions

$$x_t = \mu_x + \gamma_x x_{t-1} + \pi_t^x$$
$$y_t = \mu_y + \gamma_y y_{t-1} + \pi_t^y$$
$$z_t = \mu_z + \gamma_z z_{t-1} + \pi_t^z$$

and we store the π_t^x, π_t^y and π_t^z and estimate the first Principal Component for these residuals. From the cells in Table 2 we learn that per-whitening makes the spurious results disappear, not only for the eigenvalues and Principal Components, also for the PCR.

Illustration

What is it that we recommend to practitioners so that they can recognize non-spurious Principal Components? We recommend comparing the eigenvalues before and after pre-whitening. In case of non-spurious results, these eigenvalues should be similar.

Consider as an illustration the three annual inflation rates for France, Japan and the USA, see Franses and Janssens (2017) for data and graphs on these data and the others below. If we fit a first order autoregression to each of these variables, the estimated autoregressive coefficients obtain values of 0.931, 0.776 and 0.823, respectively. These values are all approaching 1, and we therefore should be wary for similar issues as have been observed in the simulation experiments above.

When we apply Principal Components Analysis (PCA) on the correlation matrix, we obtain for the raw data the eigenvalues 2.425, 0.446 and 0.129, and for the residuals after fitting country-specific autoregressive models of order 1, the eigenvalues 2.359, 0.418 and 0.223. Hence, in both situations there clearly is a single dominant principal component, with 0.808 and 0.786 percent of the variation explained, respectively. The weights in the first principal components are 0.610, 0.535 and 0.584 for the raw data, and 0.600, 0.553 and 0.578 for the pre-whitened data. Not only are the eigenvalues very similar, also the weights are clearly very similar.

Consider now the five annual inflation rates for the North African countries Algeria, Egypt, Libya, Morocco, and Tunisia. The first order autocorrelation are 0.772, 0.704, 0.248, 0.654, and 0.096, respectively. The first eigenvalue obtained from PCA for the raw data is 2.348 and the first principal component covers 0.470 of the total variance. The weights are 0.379, 0.421, 0.539, 0.433 and 0.448. When we fit first order autoregressions, and apply PCA to the residuals, we get a first eigenvalue of 1.870, which is associated with only 0.374 of the total variance. The weights have become 0.404, 0.213, 0.628, 0.212 and 0.594, which seem markedly different from those for the raw data. Hence, we may have found a spurious principal component here.

In Table 3, we report the estimation results for inflation in Botswana and Lesotho, two countries that are quite far away from North Africa, but for which inflation may resonate with worldwide inflation (which we assume is the first principal component for France, Japan and USA). Each first row shows that the North African principal component seems significant at close to a 5% level, while each second row shows that the World based principal component is significant at a level much less than 5%. The forecast performance of the model including the non-spurious principal component is clearly better. When we include both principal components in a single PCR, we obtain p values of 0.168 and 0.186 for the North African components, respectively. The correlation between the two principal components is only 0.335, so the low p values are not due to high correlation between these two variables. Hence, the non-spurious principal component makes the spurious component obsolete.

This illustration shows that comparing PCA outcomes for raw and pre-whitened data can be useful to diagnose non-spurious Principal Components.

Table 1: The Data Generating Process is

$$x_t = \alpha_x x_{t-1} + \varepsilon_t^x, \quad \varepsilon_t^x \sim N(0,1)$$
$$y_t = \alpha_y y_{t-1} + \varepsilon_t^y, \quad \varepsilon_t^y \sim N(0,1)$$
$$z_t = \alpha_z z_{t-1} + \varepsilon_t^z, \quad \varepsilon_t^z \sim N(0,1)$$

where it is assumed that $\alpha_x = \alpha_y = \alpha_z = \alpha$. The cells in the first panel present the average value of the first eigenvalue and the standard deviation, across 10000 replications. In the second panel, we report the frequency of significant parameters (5% level) associated with the first principal component in the PCR. There, we additionally have that $w_t = \alpha_w w_{t-1} + \varepsilon_t^w, \quad \varepsilon_t^w \sim N(0,1)$, whereas the PCR is $w_t = \mu + \rho w_{t-1} + \beta pc_{t-1} + \varepsilon_t$, with pc_{t-1} denoting the first lag of the first principal component.

	Sample size					
	50		100		500	
α						
0.5	1.288	(0.127)	1.205	(0.090)	1.091	(0.041)
0.8	1.448	(0.196)	1.328	(0.147)	1.150	(0.067)
0.9	1.567	(0.242)	1.448	(0.194)	1.219	(0.097)
0.95	1.656	(0.275)	1.568	(0.247)	1.305	(0.135)
0.99	1.786	(0.325)	1.738	(0.306)	1.572	(0.245)
α						
0.5	6.8%		5.9%		5.6%	
0.8	9.5%		6.8%		5.4%	
0.9	13.4%		9.7%		5.9%	
0.95	17.1%		13.5%		6.7%	
0.99	19.6%		18.7%		13.0%	

Table 2: The Data Generating Process is

$$x_t = \alpha_x x_{t-1} + \varepsilon_t^x, \quad \varepsilon_t^x \sim N(0,1)$$
$$y_t = \alpha_y y_{t-1} + \varepsilon_t^y, \quad \varepsilon_t^y \sim N(0,1)$$
$$z_t = \alpha_z z_{t-1} + \varepsilon_t^z, \quad \varepsilon_t^z \sim N(0,1)$$

where it is assumed that $\alpha_x = \alpha_y = \alpha_z = \alpha$. The cells in the first panel present the average value of the first eigenvalue and the standard deviation, across 10000 replications, when applied to the π_t^x, π_t^y and π_t^z , where these are the estimated residuals from

$$x_t = \mu_x + \gamma_x x_{t-1} + \pi_t^x$$
$$y_t = \mu_y + \gamma_y y_{t-1} + \pi_t^y$$
$$z_t = \mu_z + \gamma_z z_{t-1} + \pi_t^z$$

In the second panel, we report the frequency of significant parameters (5% level) associated with the first principal component in the PCR. There, we additionally have that $w_t = \alpha_w w_{t-1} + \varepsilon_t^w, \quad \varepsilon_t^w \sim N(0,1)$, whereas the PCR is $w_t = \mu + \rho w_{t-1} + \beta pc_{t-1} + \varepsilon_t$, with pc_{t-1} denoting the first lag of the first principal component.

	Sample size					
	50		100		500	
α						
0.5	1.229	(0.102)	1.160	(0.071)	1.071	(0.032)
0.8	1.230	(0.102)	1.159	(0.071)	1.070	(0.031)
0.9	1.233	(0.103)	1.159	(0.070)	1.071	(0.031)
0.95	1.233	(0.104)	1.161	(0.072)	1.071	(0.032)
0.99	1.232	(0.103)	1.161	(0.072)	1.070	(0.031)
α						
0.5	5.5%		5.0%		5.5%	
0.8	5.5%		5.4%		5.3%	
0.9	5.8%		5.5%		5.2%	
0.95	6.4%		5.4%		5.1%	
0.99	6.3%		5.6%		5.3%	

Table 3: Estimation results and evaluation of one-step-ahead forecasts, sample 1961-2015

Model I: $inflation_t = \mu + \rho\, inflation_{t-1} + \beta\, PC_{North\ Africa,t-1} + \varepsilon_t$

Model II: $inflation_t = \mu + \rho\, inflation_{t-1} + \beta\, PC_{World,t-1} + \varepsilon_t$

The data are obtained from Franses and Janssens (2017). Standard errors are given between brackets.

Country	Model	Parameter estimates		RMSPE	MAE
		ρ	β		
Botswana	I	0.536 (0118)	0.364 (0.191)	1.892	1.449
	II	0.482 (0.128)	0.498 (0.189)	1.838	1.383
Lesotho	I	-0.074 (0.141)	1.101 (0.514)	5.493	3.645
	II	-0.092 (0.138)	1.336 (0.501)	5.373	3.644

References

Ames, E. and S. Reiter (1961), Distributions of correlation coefficients in economic time series, *Journal of the American Statistical Association*, 56, 637-656.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach, *The Quarterly Journal of Economics*, 120, 387-422.

Franses, P.H. and E. Janssens (2017), Inflation in Africa, 1960-2015, Econometric Institute Report EI-2017-26, Erasmus School of Economics, https://repub.eur.nl/pub/102219

Granger, C.W.J. and P. Newbold (1974), Spurious regressions in econometrics, *Journal of Econometrics*, 2, 111-120.

Heij, C., D. van Dijk, and PJ.F, Groenen (2011), Real-time macroeconomic forecasting with leading indicators: An empirical comparison, *International Journal of Forecasting*, 27, 466-481.

Stock, J.H. and M.W. Watson (1999), Forecasting inflation, *Journal of Monetary Economics*, 44, 293-335.

Stock, J.H. and M.W. Watson (2002), Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association*, 97, 1167-1179.

Yule, G.U. (1926), Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series, *Journal of the Royal Statistical Society A* 89, 1-69.