Principal Components and Partial Least Squares analysis in data with different correlation structure

Computational Statistics, University of Bonn 2022

Krasimira Kirilova

August 23, 2022

Contents

1 Introduction **2**

2 Theorethical background of PCA and PLS **2**

2.1 Principal component analysis 2

2.2 Partial Least Squares 3

3 Simulation **4**

3.1 Setup 4

3.2 Simulation Results 5

4 Empirical Applicaiton **6**

4.1 Data set and variables 6

4.2 Application results 7

5 Conclusion **10**

References **11**

1 Introduction

Economic phenomena, both micro and macro, has always been influenced by a large number of variables. Difficulties in data collection and lack of computational power in the past imposed usage of small structural models for forecasting. Nowadays we can employ several techniques and methods to forecast economic variables using high-dimensional data. Principal component analysis(PCA) is a wildly common method for dimension reduction. The basic idea is to transform the original data in such way as to use smaller number of factors in our prediction instead of the original number of variables. Partial Least Squares(PLS) constructs the factors such that the covariance between the factors and the target variable is maximized. My goal with this project is to explore the kinds of data structure where PCA performs poorly. In the literature one of the cases studied is high-dimensional data where one variable is highly important for the target variable and the other variables are not. The many noisy variables tend to mask the signal in the leading variable when constructing the factors, leading to worse prediction. The project is organized into Theoretical background section and simulation and empirical application sections. For the lather sections I take guidance from (Groen and Kapetanios 2009) and use data set from (Stock and Watson 2009).

2 Theorethical background of PCA and PLS

As (Jolliffe 2022) writes, the first workings in PCA are from Pearson and later Hotelling in the first years of the 20th century. Decades after its creation, the developments in the method were focused on theory. The first applications to real data were done in the 60s with data sets containing 11-15 variables. Nowadays, PCA is used in variety of academic disciplines, industrial organizations and government agencies.

2.1 Principal component analysis

Given an $n \times p$ matrix X PCA creates orthogonal linear combinations of the columns of X and finds the projections which maximize the variance. The first principal component contains projections along the direction with the most variance. We derive principal components by decomposing the matrix X and deriving the eigen vectors. An $m \times n$ matrix M can be written in the form

$$M = U\Sigma V^*$$

U being an $m \times n$ unitary matrix, Σ is a diagonal $m \times n$ matrix, V is an $n \times n$ unitary matrix and V^* is the conjugate transpose of V, also unitary. In a similar manner, our data matrix X can be decomposed as

$$
\begin{aligned}
X &= U\Sigma V^* \\
X^T X &= (U\Sigma V^*)^T (U\Sigma V^*) \\
(X^T X)V &= V\Sigma^2 V^* V \\
(X^T X)V &= V\Sigma^2
\end{aligned}
$$

where V is the matrix of eigen vectors of $X^T X$ and Σ^2 is the square matrix with eigen values in the diagonal. The eigen vectors are sorted by eigen values from biggest to lowest. Now, with our eigen vectors sorted, we can project X on the new space

$$Z = XV$$

The elements of V are called principal component loadings. The columns of Z are the principal components and the elements of Z, $z_{i1}, ..., z_{in}$ are the principal component scores. Thus, the first principal component contains up to $N - 1$ of the original columns of X

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{N-1,1} X_{N-1}$$

Couple of assumptions need to be made for PCA to perform well. First, the original data needs to be standardized in order to compare the covariance between variables. Secondly, data needs to be linear, proper transformations to the original data have to be performed when this is not the case. Thirdly, the data should not have too many outliers - PCA will give more weight to highly variable features even if the variance is governed by a few outliers.

2.2 Partial Least Squares

PCA creates linear combinations that represent the original data without including the target variable. Directions that explain the original data best does not necessarily translate to directions in the data which predict the response well. PLS is an alternative dimension reduction method that takes the response variable into account. PLS identifies a set of features $Z_1, ...Z_M$ that are linear combinations of the original features and then fits a linear model using the new set of M features. Unlike PCR, PLS will not only create features that best represent the original data but also creates features that are related to the response variable. The first component in PLS is calculated by setting each principal component loading ϕ_{j1} to the coefficient from the linear regression of Y onto X_j. The highest PLS loadings correspond to the variables that are strongly related to Y. In a high-dimensional setting where one or couple of variables explains the target variable well, we'll expect PLS to outperform PCR.

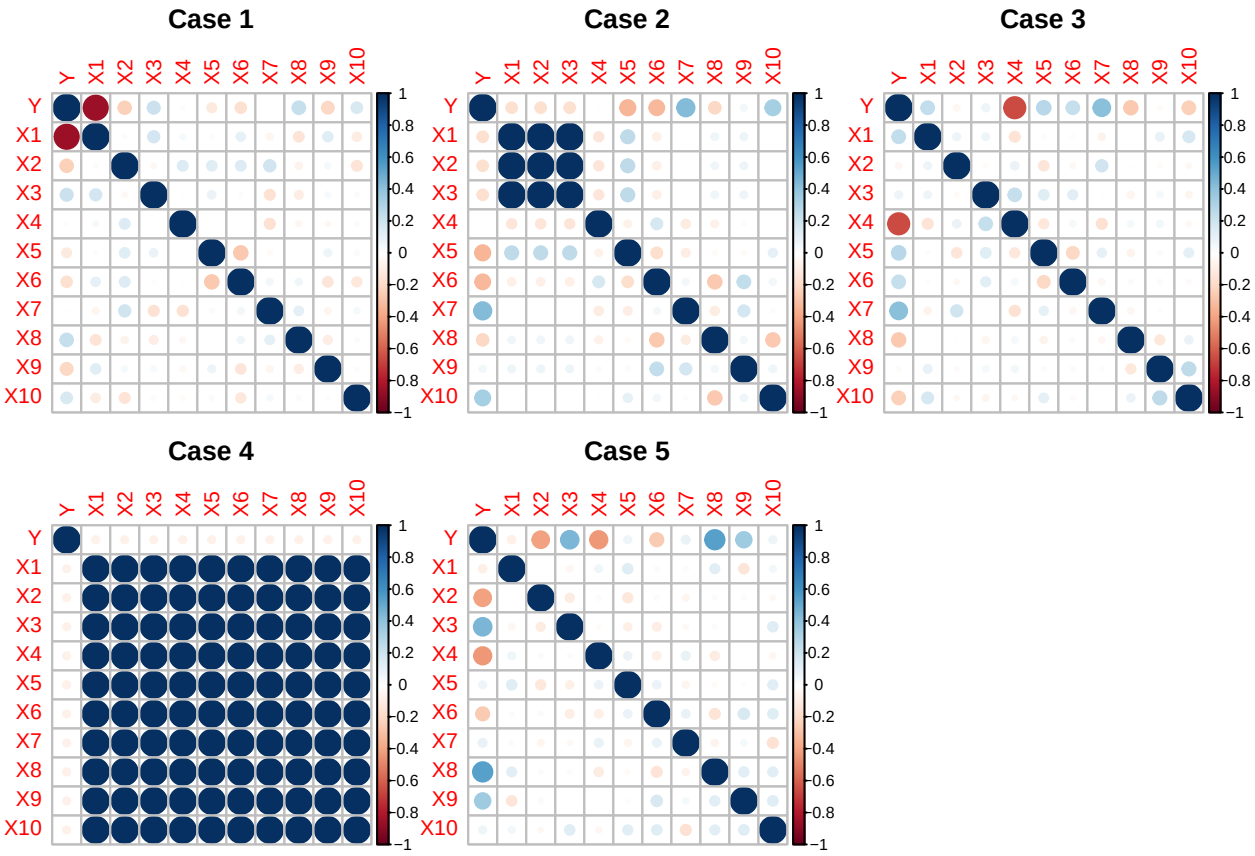
3 Simulation

The simulation study is inspired by (Groen and Kapetanios 2009). In the paper the authors compare PCR to PLS and Bayesian regression in data with different factor structures. I compare only PCR to PLS and leave Bayesian regression out.

3.1 Setup

One of the issues with principal components extracted from large number of variables is that there is no guarantee that the components will forecast the target variable well. (Boivin and Ng 2006) show that if the forecasting accuracy is driven by a certain factor, this factor can be dominated by other non-informative factors in a large data set. I simulate high dimensional data with different correlation structure under five cases:

1. Data set where one variable is highly related with the outcome variable.
2. Data set where 1/3 of the variables are highly correlated.
3. Data set where 10% of the variables are highly correlated.
4. Data set where all variables are highly correlated.
5. Data set with no high correlation between variables.



The coefficients β are generated randomly and transformed according to the cases. In the first case, I inflate the first coefficient when creating the response variable. In the other cases I inflate the coefficients proportionally to the transformed variables, where in the fifth case the proportion is 0 and the coefficients are left without transformation.

The simulations are performed using $N = 100, 200, 400, 192$ as number of observations, $p = 20, 100, 108$ as number of variables and are run 100 times, with exception of one simulation which was run for 1000 times. Models are evaluated on a training set, representing 1/2 of the sample, and predictions are made on a test set.

3.2 Simulation Results

3.2.1 Results without setting the number of components used for prediction

In the first four simulations I vary N through $N = 100, 200, 400$ and set the number of variables at $p = 20, 100$ where $p = 20$ for the first three simulations and $p = 100$ for the last one. The number of principal components used are determined by the number of components which correspond to the minimum RMSEP of the model. The change in MSE between PCR and PLSR is not quite different between the simulations, where PLSR outperforms PCR in all cases except data set with 10% of the variables are highly correlated. Both methods use 2/3 number components out of available variables in prediction on average. Poorer performance of PLSR may be attributed to bias in the regression coefficients used in constructing the principal component. Also, the small MSE in both models signals overfitting. Overall, with 2/3 variables used the difference between the methods is negligible. PLSR achieves lower MSE since it uses the response variable to construct the principal components.

3.2.2 Results with setting the number of principal components used for prediction

In next simulations, N is set to $200, 192$, p to $100, 108$ and the number of principal components used for prediction is varying $r = 1, 3, 6$. One simulation runs for 1000 times, all others for 100. $N = 192$ and $p = 108$ are chosen out of our real data set where we have 192 observations and 108 variables. When using smaller number of components again PLSR outperforms PCR in almost all cases. The only case where PCR performs better is in the case where all variables are correlated - this case is most likely to lead to overfitting compared to other cases, but even in this case when we increase the number of principal components used to six, the difference in MSE between PCR and PLSR drops to 0. PLSR performs better with less principal components used - PLSR gives way smaller MSE for one principal components used than PCR. This is not surprising given the workings of PLSR. For the most extreme cases - case where one variable is the most important for the target variable and where there is not leading variable explaining the target variable - PLSR again outperforms PCR. Mean MSE for cases 1 and 5 over simulations with different number of components used for PLSR are 40 and 73 respectively. For PCR the results are 107 mean MSE over simulations in case 1 and 73 mean MSE for case 5.

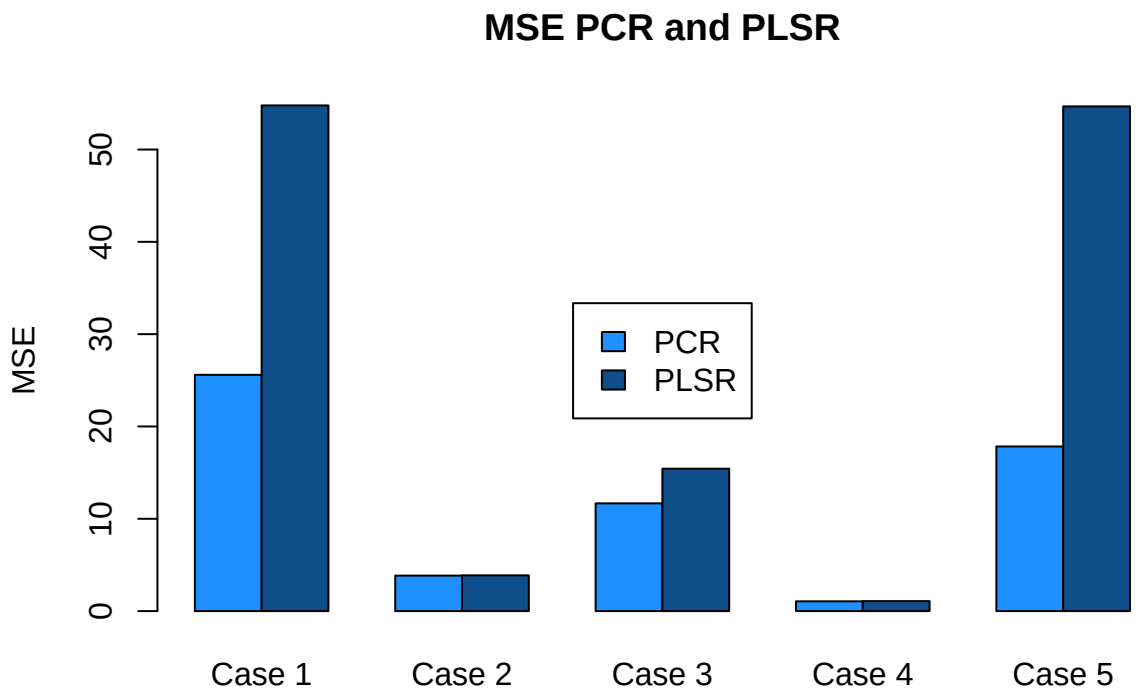


Figure 1: Simulation without setting the number of components

4 Empirical Applicaiton

To test the performance of PCR and PLSR with realistic settings, I use (Groen and Kapetanios 2009) and the data set from (Stock and Watson 2009).

4.1 Data set and variables

The data set consists of 108 macroeconomic variables in a panel going from January 1959 to December 2006. The variables used for prediction are *CPI inflation*, *industrial production*, *unemployment rate* and the *federal funds rate*. The raw data is transformed using transformation information for every variable in order to create stationary series of data. The forecasts are performed over three samples from the data: from January 1972 to December 2006, from January 1972 to December 1984 and from January 1985 to December 2006. The models are trained on sample from January 1959 to December 1971. The forecasts are updated on a window of data for every year in the sample (h = 12). The predictions from the models are tested using different number of components. For PCR, the number of components are $r = 2, 4, 6$ and for PLSR $r = 1, 2, 3$. Finally, MSE is recorder for all sub-samples and number of components and the percentage change in MSE between PCR and PLSR is

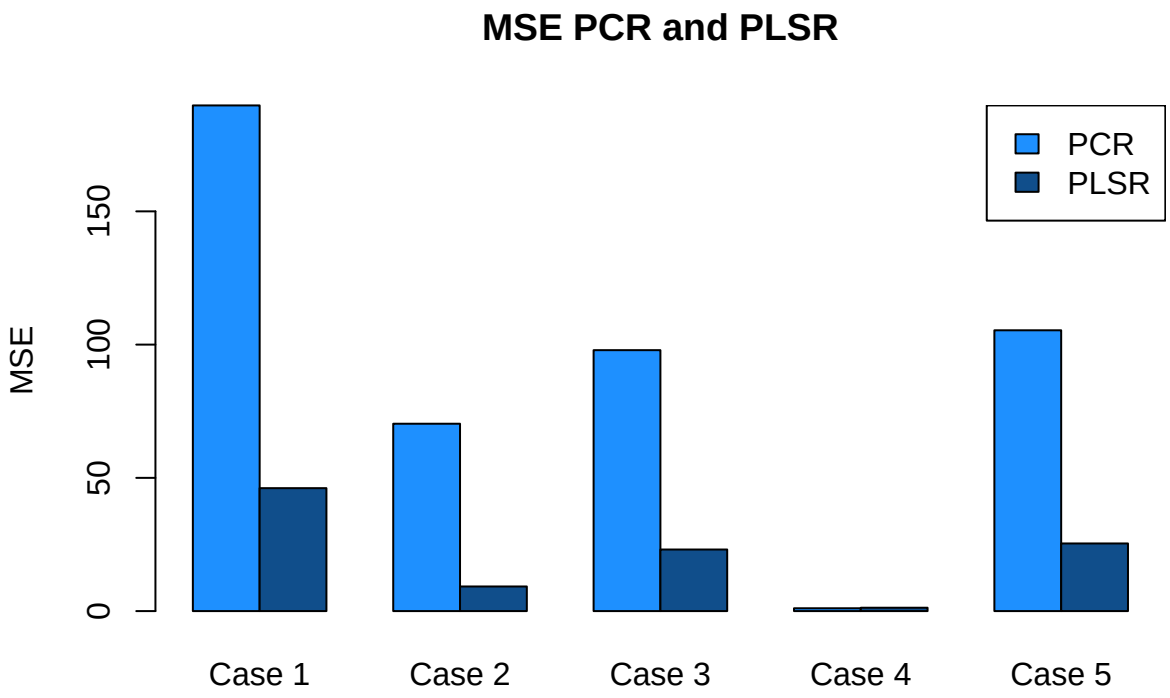


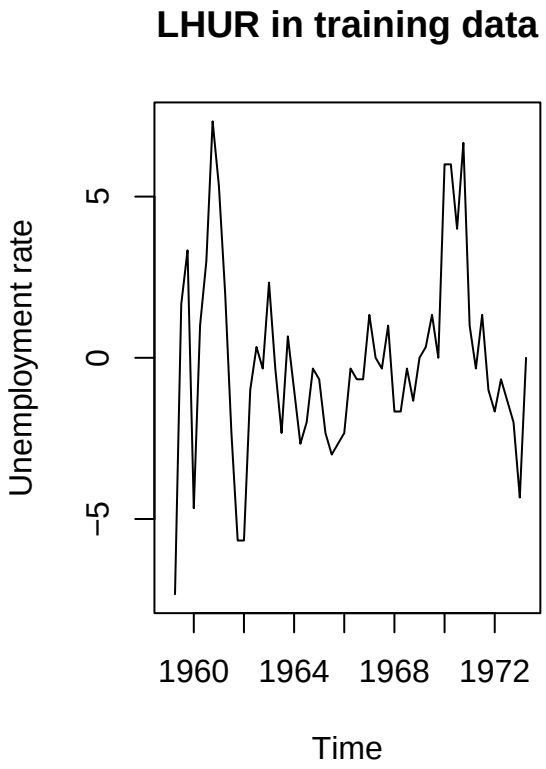
Figure 2: Simulation with setting the number of components to 6

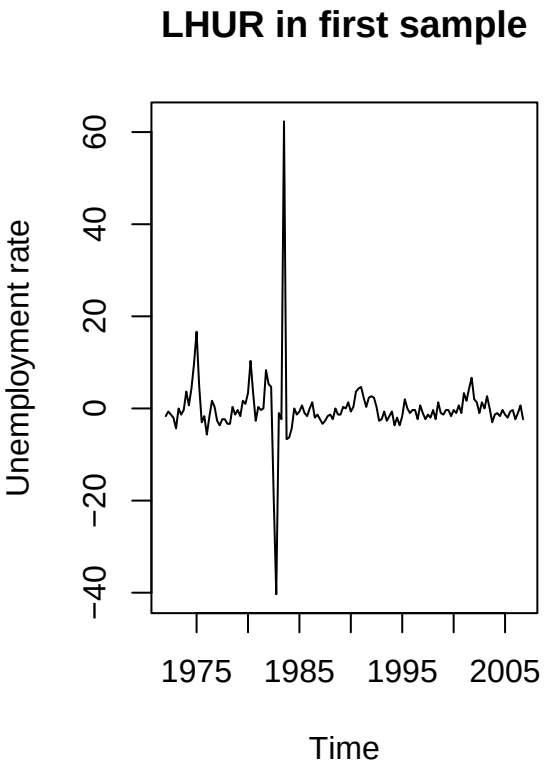
reported for every sample and variable.

4.2 Application results

I report the results using difference in MSE between PCR and PLSR in percent. Positive values indicate PLSR performed better than PCR and negative values report cases where PCR MSE was lower than PLSR MSE.
We can see the forecast performance of the methods in the three samples and for different number of components used. Cases are grouped by number of components used for each method - for PCR number of components are 2,4,6 and for PLSR number of components used are 1,2,3. Number of components used in the forecast in denoted by $r = n1, n2$ in the table output.

In the first sample, PCR outperforms PLSR for all number of components only for the unemployment rate where we see big differences in MSE for both methods. CPI inflation is predicted better by PLSR for all components. This may be the case because of substantial differences between the training set and the first test sample.





Indeed, the variation in the unemployment rate in the training set is much bigger than in the first sample. The same is true for the second sample. The dynamics in the variable are matched for the third sample only where we see that PLSR outperforms PCR in the case where small number of components(one and two respectively) are used. PLSR produces smaller MSE for CPI inflation in the first and second samples, and PCR performs better in the third sample. For CPI inflation the reasons seem to be the same as for the unemployment rate - differences between the training set and the test sets. Industrial production index is predicted better by PLSR in the first and second samples by using either one or three components. PLSR has greater MSE for industrial production index in the first two samples when we use two principal components. In the third sample, PLSR outperforms PCR only when we include three principal components in the prediction. The federal funds rate is predicted better by PLSR when we use one principal component in all samples. PCR reports smaller MSE for the federal funds rate when six principal components are used and this is true for all samples.

Table 1: First sample results

	r = 1, 2	r = 2, 4	r = 3, 6
CPI inflation	9.18 %	19.39 %	20.26 %
Industrial index	4.55 %	-12.8 %	8.53 %
Unemployment rate	-31.56 %	-628.71 %	-805.8 %
Federal funds rate	78.94 %	23.05 %	-703.42 %

Table 2: Second sample results

	r = 1, 2	r = 2, 4	r = 3, 6
CPI inflation	9.39 %	19.71 %	20.74 %
Industrial index	15.3 %	-21.21 %	-5.99 %
Unemployment rate	-101.67 %	-511.25 %	-604.65 %
Federal funds rate	61.73 %	33.96 %	-499.81 %

Table 3: Third sample results

	r = 1, 2	r = 2, 4	r = 3, 6
CPI inflation	-65.84 %	-39.11 %	-112.58 %
Industrial index	-19.14 %	-13.55 %	30.89 %
Unemployment rate	66.57 %	-828.69 %	-1596.9 %
Federal funds rate	62.5 %	-271.6 %	-745.55 %

5 Conclusion

For most applications the difference between PCR and PLSR is not substantial - both methods seem to perform the same. An important distinction between the workings of the methods arises from the underlying data structure on which they are performed. PCR is most beneficial in data setting without unique variation - all variables seem to move together. When we introduce correlation in clusters or neighborhoods of the data cloud, PLSR performs slightly better as it is able to capture the relationship between the features and the target variable. In a data set with one dominant factor explaining the target variable, PCR tends to aggregate all variables and diminish the importance of this one factor, while PLSR is able to capture the relationship and place greater predictive power on that one predictor variable - MSE for PLSR will drop substantially on the first principal component and all other components will not improve the prediction by much. Overall, the choice between PCR and PLSR lies in the structure of the data set at hand.

References

Boivin, Jean, and Serena Ng. 2006. "Are More Data Always Better for Factor Analysis?" *Journal of Econometrics* 132 (1): 169–94.

Favero, Carlo A, Massimiliano Marcellino, and Francesca Neglia. 2005. "Principal Components at Work: The Empirical Analysis of Monetary Policy with Large Data Sets." *Journal of Applied Econometrics* 20 (5): 603–20.

Fifield, Suzanne GM, David M Power, and Christopher D Sinclair. 2002. "Macroeconomic Factors and Share Returns: An Analysis Using Emerging Market Data." *International Journal of Finance & Economics* 7 (1): 51–62.

Giovannelli, Alessandro, and Tommaso Proietti. 2016. "On the Selection of Common Factors for Macroeconomic Forecasting." In *Dynamic Factor Models*. Emerald Group Publishing Limited.

Greengard, Philip, Yukun Liu, Stefan Steinerberger, and Aleh Tsyvinski. 2020. "Factor Clustering with t-SNE." *Available at SSRN 3696027*.

Groen, Jan J, and George Kapetanios. 2009. "Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting." *FRB of New York Staff Report*, no. 327.

Heij, Christiaan, Dick van Dijk, and Patrick JF Groenen. 2008. "Macroeconomic Forecasting with Matched Principal Components." *International Journal of Forecasting* 24 (1): 87–100.

Jolliffe, Ian. 2022. "A 50-Year Personal Journey Through Time with Principal Component Analysis." *Journal of Multivariate Analysis* 188: 104820.

Marcellino, Massimiliano, James H Stock, and Mark W Watson. 2003. "Macroeconomic Forecasting in the Euro Area: Country Specific Versus Area-Wide Information." *European Economic Review* 47 (1): 1–18.

Stock, James H, and Mark Watson. 2009. "Forecasting in Dynamic Factor Models Subject to Structural Instability." *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry* 173: 205.

Stock, James H, and Mark W Watson. 2012. "Generalized Shrinkage Methods for Forecasting Using Many Predictors." *Journal of Business & Economic Statistics* 30 (4): 481–93.