

On the use of partial least squares regression for forecasting large sets of cointegrated time series

Gianluca Cubadda

Universita' di Roma 'Tor Vergata', e-mail: gianluca.cubadda@uniroma2.it

Barbara Guardabascio

Universita' di Roma 'Tor Vergata', e-mail: barbara.guardabascio@uniroma2.it

Abstract: This paper proposes a methodology to forecast cointegrated time series using many predictors. In particular, we show that partial least squares can be used to estimate single-equation models that take into account of possible long-run relations among the predicted variable and the predictors. Based on Helland (1990), and Helland and Almqvist (1994), we discuss the conditions under which partial least squares regression provides a consistent estimate of the conditional expected value of the predicted variable. Finally, we apply the proposed methodology to a well-known dataset of US macroeconomic time series (Stock and Watson, 2005). The empirical findings suggest that the new method improves over existing approaches to data-rich forecasting, particularly when the forecasting horizon becomes larger.

Keywords: Partial Least Squares, Error Correction Model, Forecast, large data - sets

1. The model

Let us consider the vector error-correction model of an n -vector of time series z_t

$$\Gamma(L)\Delta z_t = \alpha\gamma'z_{t-1} + \varepsilon_t, \quad (1)$$

where $\Delta = (1 - L)$, $\Gamma(L) = I_n - \sum_{i=1}^p \Gamma_i L^i$, α and γ are full-column rank $n \times r$ ($r < n$) matrices such that $\alpha'_{\perp} \Gamma(1) \gamma_{\perp}$ has full rank, ε_t are i.i.d. $N_n(0, \Sigma_{\varepsilon\varepsilon})$, and, for simplicity, no deterministic elements are introduced. The process z_t is then cointegrated of order (1,1) and the matrix γ span the cointegrating space, see e.g. Johansen (1996).

It follows that each of series z_t is generated by a model with the following form:

$$y_t = \beta'x_{t-1} + \epsilon_t, \quad t = 1, \dots, T \quad (2)$$

where y_t is a generic element of Δz_t , ϵ_t is the corresponding element of ε_t , and $x'_t = [z'_t\gamma, \Delta z'_t, \dots, \Delta z'_{t-p+1}]$.

Suppose that $N = pn + r$, the number of elements of β , is too large to estimate model (2) by ordinary least squares. In order to reduce the number of parameters in (2), we assume that

$$\sigma_{xy} = \Upsilon_q \xi, \quad (3)$$

where $\sigma_{xy} = E(x_{t-1}y_t)$, Υ_q is a matrix formed by q eigenvectors of $\Sigma_{xx} = E(x_t x'_t)$, and ξ is a vector with all the q elements different from zero. Notice that equation (3) implies

$$\beta = \Upsilon_q \Lambda_q^{-1} \xi$$

where Λ_q is the diagonal eigenvalue matrix associated with Υ_q . Hence, model (2) has the following factor structure:

$$y_t = \xi' f_{t-1} + \epsilon_t,$$

where $f_t = \Lambda_q^{-1} \Upsilon_q' x_t$. Notice that $y_{t+h|t} = E(y_{t+h}|f_t)$ is a linear combination of the factors f_t for $h > 0$.

2. Estimation of the factors

Based on Helland (1990) and Helland and Almoy (1994), we argue that, under condition (3), partial least squares (PLS) regression provides a consistent estimator of β . Indeed, at the population level the coefficient matrix of the q PLS factors $f_{t-1}^{PLS} = \Omega_q' x_{t-1}$ is obtained (up to a non-singular linear transformation) by the recursive relation.

$$\omega_{i+1} = \sigma_{xy} - \Sigma_{xx} \Omega_i (\Omega_i' \Sigma_{xx} \Omega_i)^{-1} \Omega_i' \sigma_{xy} \quad (4)$$

where $\Omega_i = (\omega_1, \dots, \omega_i)$ for $i > 0$, and $\omega_1 = \sigma_{xy}$. It follows by induction that Ω_q lies in the space spanned by

$$(\sigma_{xy}, \Sigma_{xx} \sigma_{xy}, \dots, \Sigma_{xx}^{q-1} \sigma_{xy}),$$

from which it is easy to see that condition (3) implies $\omega_{q+1} = 0$ and

$$\beta_{PLS} \equiv \Sigma_{xx} \Omega_q (\Omega_q' \Sigma_{xx} \Omega_q)^{-1} \Omega_q' \sigma_{xy} = \Sigma_{xx} \sigma_{xy}. \quad (5)$$

Since the PLS algorithm by Wold (1985) builds q factors from x_{t-1} such that their coefficient matrix lies in the space spanned by

$$(\hat{\sigma}_{xy}, \hat{\Sigma}_{xx} \hat{\sigma}_{xy}, \dots, \hat{\Sigma}_{xx}^q \hat{\sigma}_{xy}),$$

where $\hat{\sigma}_{xy}$ and $\hat{\Sigma}_{xx}$ are, respectively, the natural estimators of σ_{xy} and Σ_{xx} , we see that $\hat{\beta}_{PLS}$, i.e. the sample analogue of (5), is a consistent estimator of β . Remarkably, the original Wold's PLS algorithm computes the factors $\hat{f}_{t-1}^{PLS}$ without requiring any covariance matrix inversion. Hence, this method is very attractive when N is large compared to the sample size T . A problem arises about the estimation of the cointegrating matrix γ . Since r is unknown and we can not resort to a maximum likelihood approach (Johansen, 1996) due to the large predictor number N , we use the principal components of z_{t-1} in place of $\gamma' z_{t-1}$ in model (2). The rationale for this approach is that the eigenvectors associated with the r smallest eigenvalue will be a super-consistent estimator of the cointegrating vectors γ whereas the other $(n - r)$ principal components are $I(1)$, hence they are asymptotically uncorrelated with the $I(0)$ predicted variable y_t . We notice that Gröen and Kapetanios (2008) have recently proposed to use PLS for macroeconomic forecasting in a data-rich environment. However, they do not explicitly consider the existence of long-run relations among the variables of interest. Moreover, their method requires that both $N \rightarrow \infty$ and $T \rightarrow \infty$ to ensure consistency of the estimated PLS factors, whereas only T need to diverge in our approach.

3. A Forecasting Exercise

We employ the same dataset as Stock and Watson (2005), which contains 131 US monthly time series covering a broad range of economic categories such as industrial production,

income, unemployment, employment, wages, stock prices, and consumer and producer prices. The time span is from 1959.01 through to 2003.12. Variables are transformed to achieve stationarity. Moreover, we compute the principal components of the 103 time series that appear to be $I(1)$, possibly after differencing. The first lag of these principal components are then included in the predictor set in place of the unknown error correction terms. Finally, the predicted variables are Industrial Production (IP), Employment (EMP), Federal Funds Rate (FYFF), and Consumer Price Index (CPI). Following Stock, Watson 2002 and De Mol, Giannone 2008 for each of these four variables, h -step ahead forecasts for $h = 1, 3, 6, 12$ are obtained as

$$\begin{aligned}\Delta_h \widehat{z}_{i,t+h} &= \widehat{\xi}_{i,h}' \widehat{f}_t^{PLS} \\ \Delta_h &= (1 - L^h) \\ i &= IP, EMP, FYFF, CPI\end{aligned}$$

where both the PLS factors and their loadings are estimated using a rolling window of 299 observations. In particular, the loadings are computed by GLS, allowing for both heteroskedasticity and autocorrelation of order $(h - 1)$. The number of PLS factors, q , and the number, p , of lags of Δz_t to be included in x_{t-1} , p , are fixed by minimizing the 6-step head mean square forecast error (MSFE) that is computed using the training sample 1959.01-1979.12 and the validation sample 1980.01-1984.12.

In order to check out the best number of long-run relations r to be considered in the model, we apply a second training procedure. First of all, we reorder the principal components according to magnitude of their PLS regressions coefficients, then we run n PLS regressions such that the i -th model includes i reordered principal components for $i = 1, \dots, n$. Finally, we choose the model that achieves the lowest 6-step head MSFE given the values of p, q previously selected. Hence, the forecasting exercise is carried out for the period 1985.01-2002.12. In Table 1 we report the MSFE relative to the naive random walk forecast for both the PLS and the Bayesian regression (BR) forecasts obtained using a rolling window of 120 observations (De Mol *et al.*, 2008).

Table 1: Relative MSFE's of PLS and BR

RMSFE of PLS forecasts					RMSFE of BR forecasts				
Variables	IP	EMP	FYFF	CPI	Variables	IP	EMP	FYFF	CPI
N. of Lags	1	1	1	1					
N. PLS Components	2	2	1	3					
N. PCA Components	22	19	4	13					
h= 1	0.85	0.50	0.63	0.33	h= 1	0.89	0.48	0.77	0.82
h= 3	0.71	0.36	0.57	0.06	h= 3	0.81	0.36	0.62	0.95
h= 6	0.79	0.43	0.59	0.01	h= 6	0.89	0.50	0.62	0.92
h= 12	0.87	0.55	0.62	0.01	h= 12	1.05	0.70	0.70	0.91

The comparison reveals that the PLS forecasts are often more accurate than those obtained by BR, particularly when the forecasting horizon is large. This outcome is likely due to the inclusion of the error correction terms in the predictor set. Anyway, PLS sometime improves over BR even for short forecasting horizon, see e.g. the CPI forecasts. In the few cases when PLS is outperformed by BR, the former method is still a close competitor.

References

- [1] De Mol, C., Giannone, D. (2008), Forecasting using a large number of predictors: is Bayesian regression a valid alternative to principal components?, *Journal of Econometrics*, 3066.
- [2] Gröen, J.J., and G. Kapetanios (2008), Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting, *Federal Reserve Bank of New York Staff Reports*, 327.
- [3] Helland, S. (1990), Partial Least Squares Regression and Statistical Models, *Scandinavian Journal of Statistics*, 17, 97 - 114 (1990).
- [4] Helland, S, and T. Almoy (1994), Comparison of prediction methods when only a few components are relevant, *Journal of the American Statistical Association*, 89, 583-591.
- [5] Johansen, S. (1996), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models* (Oxford University Press: Oxford).
- [6] Stock, J. H., and M. W. Watson (2005), An Empirical Comparison Of Methods For Forecasting Using Many Predictors, Manuscript, Princeton University
- [7] Wold, H. (1985), Partial least squares, in S. Kotz and N.L. Johnson (eds), *Encyclopedia of the Statistical Sciences*, John Wiley & Sons, 6, 581–591.