UNIVERSIDAD CARLOS III DE MADRID
working papers

Working Paper 12-22
Statistics and Econometrics Series 16
August 2012

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

Sparse Partial Least Squares in Time Series for Macroeconomic Forecasting

Julieta Fuentes[1], Pilar Poncela[2], Julio Rodríguez[3]

Abstract

Factor models have been applied extensively for forecasting when high dimensional datasets are available. In this case, the number of variables can be very large. For instance, usual dynamic factor models in central banks handle over 100 variables. However, there is a growing body of the literature that indicates that more variables do not necessarily lead to estimated factors with lower uncertainty or better forecasting results. This paper investigates the usefulness of partial least squares techniques, that take into account the variable to be forecasted when reducing the dimension of the problem from a large number of variables to a smaller number of factors. We propose different approaches of dynamic sparse partial least squares as a means of improving forecast efficiency by simultaneously taking into account the variable forecasted while forming an informative subset of predictors, instead of using all the available ones to extract the factors. We use the well-known Stock and Watson database to check the forecasting performance of our approach. The proposed dynamic sparse models show a good performance in improving the efficiency compared to widely used factor methods in macroeconomic forecasting.

Keywords: Factor Models, Forecasting, Large Datasets, Partial Least Squares, Sparsity, Variable Selection

[1] Julieta Fuentes, Universidad Carlos III de Madrid and Banco Central de Reserva de El Salvador, Alameda Juan Pablo II, San Salvador, e-mail: julieta.fuentes@alumnos.uc3m.es; julieta.fuentes@bcr.gob.sv. Corresponding autor

[2] Pilar Poncela, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049, Madrid, e-mail: pilar.poncela@uam.es

[3] Julio Rodríguez, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049, Madrid, e-mail: jr.puerta@uam.es

Acknowledgements: Pilar Poncela and Julio Rodríguez acknowledge financial support from the Spanish Ministry of Education, contract grant ECO2009-10287.

INTRODUCTION

The availability of large data sets in many fields provides the possibility of improving the forecast performance of the target variables. Several approaches have been developed to deal with problems that are ill-posed due to the high dimensionality and multicollinearity of the information sets. Some have proved to be useful in the empirical analysis in areas such as finance and economics (see, for instance, Stock and Watson, 2002a, b; De Mol et al., 2008; Matheson, 2006; among many others). In particular, factor models have been applied in two steps to solve this problem: first, the information contained in a large data set of predictors is reduced through factor analysis and, second, the estimated factors obtained in the previous step are regressed over the target variable.

The main idea behind factor models and related techniques is to reduce the dimension of the subspace spanned by the predictors. There are several ways of estimating the common factors. Principal components (PC) constitute the standard and most widely used method, be they static or dynamic (Stock and Watson, 2002a and Forni et al, 2000 and 2005). They provide a consistent estimator for the factors contained in the predictors (called X) in large approximate dynamic factor models, when both the dimension of the cross section N and the sample size T go to infinity. Recent surveys can be found in Stock and Watson (2006, 2010) and Bai and Ng (2008).

One criticism of factor methods, such as PC, is that the estimated factors do not depend directly on the prediction purpose and thus the data considered in the panel might not have any predictive power over the target variable. That is, the process of reducing the dimension among the predictors is not related to the forecasted goal. In a static context, Partial Least Squares (PLS from now on) was introduced by Wold (1966) who considered the goal of prediction while reducing the dimension of the subspace spanned by the predictors. This was a valid alternative even in the extreme case where the number of predictors, N, was larger than the sample size, T, and/or in cases of severe multicollinearity. In this paper, we revisit PLS and discuss its implementation, both static and dynamic, for time series in accordance with the properties of the data considered. Groen and Kapetanios (2008) and Wang (2008) are early attempts in this line.

Furthermore, there is growing evidence that the number, like the quality of the variables included in the dataset, is relevant for the estimation process (see, for instance, Bai and Ng, 2002; Boivin and Ng, 2006; Watson, 2000; Stock and Watson, 2007a; Eickmeier and Ng, 2011; among others). Since the factor space being estimated is a function of the chosen panel of the predictor variables, the information content on the data is crucial to improve forecasting accuracy. Hence, instead of using all available predictors, we focus on the choice of a useful or informative subset of them to extract the latent variables and forecast a specific target variable. Notice that the "factors", or to be more precise the unobserved common components, do not need to load in a great number of variables. On the contrary, the factor loadings can be zero for many predictors if they do not have enough informative content about the target variable. Bai and Ng (2008) is a first approximation to overcome this problem in the context of PC.

We introduce Sparse Partial Least Squares (SPLS) into the economic analysis, a technique that, besides taking into account the response variable for the component estimation, allows a variable selection process to be performed to construct a factor-forming subset. The SPLS method has been used in chemometrics in a static context (Chun and Keles, 2010). We also propose its dynamic extension as in the case of PLS.

Another feature of our proposal is that for stable relationships, it allows the best predictors to be selected within a large data set, so it can be used as an exploratory tool. For unstable relationships, the re-estimation of the model allows the selection of the best predictors within the data set based on their predictive content by monitoring the variables that go in/out of the model.

We use the Stock and Watson data base in order to perform an empirical comparison of the forecasting performance of the PLS and SPLS methods to those widely used nowadays as principal components and targeted predictors. We focus our attention on forecasting inflation motivated by the reported difficulty to improve its performance, which is due in part to the changes in the inflation process and therefore in the instability of its predictive relationships (Stock and Watson, 2007a,b). The main findings confirm that there is some room for refinement in the factor forecasting methodology.

The paper is organized as follows. Section 2 presents the forecasting framework and briefly describes the factor methods most frequently used. Section 3 discusses PLS in a dynamic context. Section 4 reviews some regularization methods already used in economics and introduces the sparse version of PLS (SPLS). We discuss its implementation, both static and dynamic. Section 5 presents the empirical application and provides the forecasting comparison for several horizons. Section 6 concludes.

2. FORECASTING FRAMEWORK

Our goal is to forecast y_{t+h} given the available information of the target up to time t, as well as from many other predictors, that we denote as X_t, and their lags. Since X_t can incorporate a large number of predictors, we would like to extract the information that is valuable for forecasting y_{t+h} in a parsimonious way. If the common information in X_t coincides with the useful information for forecasting y_{t+h}, we can use factor techniques to extract it. We use the term "factor" in a broad sense, meaning the unobserved component or signal that might be common to several variables (although not necessarily to many of them).

The forecasting model is specified and estimated as a linear projection of a h-step ahead transformed variable y_{t+h} onto t dated predictors. The predictors are the estimated factors, their lags and lags of the variable to be forecasted. That is,

$$y_{t+h} = \mu + \phi'(L) y_t + \beta'(L) \hat{F}_t + \eta_{t+h} \qquad (1)$$

where y_{t+h} is the variable to be forecasted at period t+h as a function of its own lags $\phi(L)$ y_t and of the factors and their lags estimated in the previous step $\beta(L)\,\hat{F}_t$. The h-step ahead prediction error is denoted by η_{t+h}. The factor methods differ both in the way in which the factors are extracted and in the way in which the projection of the common component is made.

In what follows we review the main approximations that have appeared in the literature to estimate the unobserved factors for large N.

2.1 Factor Models

In factor models variables are represented as the sum of two mutually orthogonal unobservable components: the common and the idiosyncratic components. The common component is driven by a small number of factors common to all variables in the model and the idiosyncratic one is driven by variable specific shocks. Assume that the data admit a factor model representation given by:

$$X_t = \Lambda F_t + \varepsilon_t \qquad (2)$$

where $X_t = (X_{1t}, \ldots, X_{Nt})$ is a N-dimensional vector of time series observed at t of candidate predictors, F_t is a k x 1 vector of common factors, $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_k]$ is a N x k matrix of factor loadings and ε_t is a Nx1 vector of idiosyncratic disturbances. The factors and the idiosyncratic disturbances are assumed to be uncorrelated at all leads and lags, that is, $E(f_t \varepsilon_{is})=0$ for all i, s. A factor model with orthogonal idiosyncratic elements is called a strict factor model while an approximate factor model relaxes this assumption and allows a limited amount of correlation among the idiosyncratic terms.

The maximum likelihood estimation of the dynamic factor model for small models (N is considered small and finite) has been known for a long time in the literature (see, for instance, Geweke, 1977; Geweke and Singleton, 1981 and Engle and Watson, 1981; for early contributions to this literature). However, recent results given by Jungbacker and Koopman (2008) and Doz et al., (2012) allow the estimation of large N dynamic factor models by maximum likelihood using the state space framework and Kalman filter techniques for large N models. Moreover, Doz et al., (2012) show that the common factor estimates are consistent even though there is weak cross correlation in the error term not taken into account in the estimation procedure.

In the case of approximate factor models with large N and stationary factors, there are several estimation approaches based on principal components, such as static principal components and extensions as weighted principal components.

2.2 Principal Components (PC)

Stock and Watson (2002a) model the covariablility of a large number of predictor series (N) in terms of a small number of unobserved latent factors, and they build the forecasts using a linear regression between these estimated latent factors and the variable to forecast. Assume that X_t admits a factor model representation as in (2). The estimation of the factors is performed using the first k principal components of $\{X_t\}_{t=1}^{T}$, which are obtained by solving the following minimization problem in $\tilde{\lambda}_i$ and $\tilde{F}_t$

$$\min V = \min \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \hat{X}_{it})^2 = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \tilde{\lambda}_i \tilde{F}_t)^2. \tag{3}$$

The solution of (3) provides the approximation with minimum mean square error for the X matrix. The problem is usually rewritten as the maximization of $N\,tr(\tilde{\Lambda}' X' X \tilde{\Lambda})$ subject to the identification restriction $\tilde{\Lambda}'\tilde{\Lambda}=I_r$ where tr (·) denotes the matrix trace. The objective is then to find the maximizer vectors ($\hat{\Lambda}$) of the diagonal sum of $\tilde{\Lambda}' \Sigma_{xx} \tilde{\Lambda}$, which is solved setting $\hat{\Lambda}$ equal to the eigenvectors of X′X corresponding to its k largest eigenvalues. Then, the estimator of the factor is constructed as $\hat{F}_t = \hat{\Lambda}' X_t$, the vector consisting of the first k principal components of X_t.

The model proposed by Stock and Watson (2002a) is relatively simple to apply. Boivin and Ng (2005) showed that factor model based on static principal components is quite robust to misspecification since very few auxiliary parameters have to be specified. However, it does not exploit the dynamics of the common factors.

Forni et al. (2005) proposed a different scheme for the estimation of the factors, weighted principal components. They estimated the common factors based on generalized principal components (GPC) in which observations are weighted according to their signal to noise ratio. The estimation is carried out in the frequency domain and the factors can load dynamically onto the observed predictors in the so called generalized dynamic factor model. Other weighting schemes have been proposed in the literature, such as the one by Inoue and Kilian (2008), which produce the weights by bootstrapping forecast based on pretest model selection.

3. PARTIAL LEAST SQUARES (PLS)

For the relationship between the target variable and the set of predictors when reducing the dimension, we use PLS. Partial Least Squares is a dimension reduction technique originally proposed by Wold (1966). PLS constructs a scheme for extracting orthogonal unobserved components based on the covariance between the predictors and the dependent or forecasting variable ($X'Y$). The components are estimated from successive iterations of the optimization problem. The factors or components could be obtained from the eigenvalue decomposition of the matrix:

$$M = X'YY'X. \tag{4}$$

The first PLS factor $\hat{f}_{jt}^{PLS}$ is determined by a linear combination of the predictor variables in X and the first eigenvector of M. To find the second PLS factor, the eigenvalue decomposition is performed on the residuals of the simple regressions of the target variable Y as well as on each of the predictors in X over the first PLS component. These residuals contain the information that is unexplained by the first PLS factor. The process is repeated until the last factor has been extracted.

It is important to note that Partial Least Squares seeks directions that have high variance and high correlation with the forecasting variable, in contrast to principal components regression, which focuses only on high variance (Hastie et al., 2008).

PLS has been implemented in a static way while its implementation taking into account the dynamic behavior of the target is scarce (see, Groen and Kapetanios, 2008 and Eickmeier and Ng, 2011). We review the basic static application (static approach) and revisit and discuss how PLS can be applied to time series (dynamic approaches). We examine several types of approximations that account for the dynamics of the time series from alternative perspectives. The approaches differ in the set of predictors, the definition of the target adopted when extracting the factors and the estimation procedure.

To define the forecasting model, consider the two equations:

$$y_{t+h} = \beta'(L)Z_t + \phi(L)' y_t + u_{t+h} \tag{5}$$

and

$$Z_t = WX_t. \tag{6}$$

Equation (5) is our forecasting equation to produce the h-step ahead forecasts of the target variable, y. Forecasts are built as the sum of two components: their own dynamics collected in the term $\phi'(L)y_t$; and the influence of the unobserved common components $\beta'(L)Z_t$. To highlight the difference between principal component regression and PLS, we denote the PLS components by $Z_t = \hat{f}_t^{PLS}$. The h-step ahead prediction error is denoted by u_{t+h}. Equation (6) expresses that the unobserved common components are formed as linear combinations of the candidate predictors through the weighting matrix W, where W is Nxk.

The key issue is how to estimate the unobserved components Z_t taking into account that we are dealing with time series. In what follows we discuss several static and dynamic possibilities.

Static Approach

a. The factors are extracted by applying PLS between the target variable (Y_{t+h}) and the original set of predictors (X). The lags of the target variable are included in the forecasting equation (5), while they are not taken into account when forming the unobserved common components Z_t. The M matrix is given by $M= X'Y_hY_h'X$, where $Y_h=(y_{h+1},\ldots,, y_{T+h})$ is the vector containing the target h periods ahead.

Dynamic approaches (DPLS)

b. The factors are based on applying PLS between an expanded set of predictors (X_e), enlarged with lags of the target variable, and the target variable (Y_h). The forecasting equation (5) does not include lags of the target which are added as additional predictors in the linear combinations formed in (6).

c. The factors are based on applying PLS between the original set of predictors (X) and the residuals from an AR (p) process fitted to the target variable (Y_h). This can be done in a two step estimation procedure or in an iterative estimation algorithm. The lags of the target variable are included in the forecasting equation (5).

To illustrate the main ideas and see the potential advantages and disadvantages of each of the possibilities of applying PLS to time series data, we consider the simple case when the number of unobserved "factors" is k=1, the number of predictors is N=2 and the number of lags of the variable to be forecasted is just 1; so the AR filter in equation (5) is just $\phi(L)=\phi$, and we only need to include y_t. For the static approach a, the h period ahead forecast is generated by the two step estimation of the following equations:

$$y_{t+h} = \beta_1 Z_t + \phi y_t + u_{t+h} \tag{7}$$

$$Z_t = w_1 x_{1t} + w_2 x_{2t} \tag{8}$$

where $Z_t = \hat{f}_t^{PLS}$ and w_i, i=1,2, are the weights assigned to each one of the predictor variables in the PLS component.

In a first step, the direction vector w is found by solving the following optimization problem:

$$w = \operatorname{argmax}_{w} \; w' X_t' Y_{t+h} Y_{t+h}' X_t w \quad \text{subject to} \quad w'w = 1 \tag{9}$$

with w=(w_1, …, w_r)' which leads to the following objective function for the case r=2

$$\max_{(w_1 w_2)} w_1^2 [\sum_{t=1}^{T} x_{1t} y_{t+h}]^2 + 2 w_2 w_2 [\sum_{t=1}^{T} x_{1t} y_{t+h}][\sum_{t=1}^{T} x_{2t} y_{t+h}] + w_2^2 [\sum_{t=1}^{T} x_{2t} y_{t+h}]^2 + \lambda (w_1^2 + w_2^2 - 1)$$

where $[\sum_{t=1}^{T} x_{it} y_{t+h}]$ is T times the covariance between each predictor in X and Y_{t+h}. Solving the previous problem, we obtain that in the first PLS component the direction vector w is a function of the covariances between each of the predictors (X_t) and the target variable (Y_{t+h}):

$$w_r = \frac{\sum_{t=1}^{T} x_{it} y_{t+h}}{\sqrt{[\sum_{t=1}^{T} x_{1t} y_{t+h}]^2 + [\sum_{t=1}^{T} x_{2t} y_{t+h}]^2}} \quad for\ r = 1, 2. \tag{10}$$

In a second step, once the factor $\hat{f}_t^{PLS} = Z_t$ has been estimated, it enters equation (7) to serve as a reduced set of explanatory variables. The dynamic relationships of the target variable are captured directly through the inclusion of its own lags as explanatory variables in the forecasting equation.

In the dynamic approach b, the model set up is as follows:

$$y_{t+h} = \beta_1 Z_t + u_{t+h} \tag{11}$$

$$Z_t = w_1 x_{1t} + w_2 x_{2t} + w_3 y_t . \tag{12}$$

In this case, instead of incorporating the lags of the target variable (Y_t) as regressors in the forecasting equation, they are included as additional predictors in X. In our simple illustration, the expanded data set contains three predictor variables, where $x_{3t}=y_t$. The direction vectors are estimated by solving the optimization problem (9), for r=3:

$$w_r = \frac{\sum_{t=1}^{T} x_{rt} y_{t+h}}{\sqrt{[\sum_{t=1}^{T} x_{1t} y_{t+h}]^2 + [\sum_{t=1}^{T} x_{2t} y_{t+h}]^2 + [\sum_{t=1}^{T} y_t y_{t+h}]^2}} \quad for \;\; r = 1, 2, 3. \qquad (13)$$

Notice that if PLS assigns a weight to all the variables included in the data set, the AR(p) process associated to the target variable (AR(1) in this simple setup) will attenuate its participation as $N \rightarrow \infty$. Then, if the AR(p) process is relevant for explaining the target variable, as is the case for macroeconomic variables, this approach could have a poor performance relative to the "static approach", where the AR(p) process is included directly in the forecasting equation.

Approach c proposed an alternative way to integrate the dynamic relationship in the factor estimation that consists in isolating the effect of the AR(p) process before the PLS estimation. The forecasting framework can be expressed as in (7) and (8), but the optimization problem (9) is modified as follows:

$$w_r = \operatorname{argmax}_w \; w' X_t' Y Y' X_t w \quad \text{subject to} \quad w' w = 1 \qquad (14)$$

where Y= [Y_{t+h}-ϕY_t]. In our example, the problem can be stated as:

$$Max_{w_k} = [w_1[\sum_{t=1}^{T} x_{1t}(y_{t+h} - \phi y_t)] + w_2[\sum_{t=1}^{T} x_{2t}(y_{t+h} - \phi y_t)]]^2 + \lambda(w_1^2 + w_2^2 - 1).$$

The estimated direction vectors for this alternative method preserve the same structure as the previous ones but depend on the AR(1) coefficient:

$$w_r = \frac{\sum_{t=1}^{T} x_{it}(y_{t+h} - \phi y_t)}{\sqrt{[\sum_{t=1}^{T} x_{1t}(y_{t+h} - \phi y_t)]^2 + [\sum_{t=1}^{T} x_{2t}(y_{t+h} - \phi y_t)]^2}} \quad for\ r = 1, 2 \qquad (15)$$

where ϕ is the autoregressive coefficient of the AR(1) model that captures the target variable's own dynamics, before the PLS estimation.

4. SPARSE METHODS

There is a growing body of literature which suggests that the selection of relevant variables from a large feasible set is needed to improve forecast efficiency. Bai and Ng (2008) proposed forecasting economic series using a reduced set of informative variables named targeted predictors (TP). The authors combine a variable selection process with principal components estimation. Two types of threshold rules (hard and soft thresholds) are introduced in order to take into account the relation between the whole dataset and the variable of interest.

The hard threshold procedure is based on a statistical test to screen the variables from the dataset considered. In this case, the targeted predictors are selected as follows:

(a) A regression is performed between the variable to be forecasted (y_{t+h}) and each predictor variable; one constant and four lags of y_t are also included in the application.

(b) A threshold significance level α is set.

(c) A smaller set of predictors $^{*}_{\propto}$, whose t ratio defined as $|\hat{\beta}_i / se(\hat{\beta}_i)|$ exceeds the predefined threshold, are selected as targeted predictors.

(d) The factors are extracted by principal components from the reduced dataset ($^{*}_{\propto}$).

(e) The forecast equation based on the previous extracted factors is estimated.

The soft threshold procedure intends to solve some drawbacks of the hard threshold method such as the sensitiveness of the estimation to small changes in the data due to the discreteness of the decision rule and the fact that the information within predictors is not considered in the selection process. As an alternative procedure to hard thresholding, the authors proposed using penalized regressions. In their empirical analysis they employed least angle regression (LARS) to select a subset of variables before performing the principal component analysis.

In order to overcome simultaneously the two drawbacks pointed out for extracting the "factors" (not taking into account the forecasting goal and too much uncertainty because weight is given to all predictors) Chun and Keles (2010) propose the SPLS static formulation in the context of biology. The SPLS approach imposes an additional constraint (λ) on the PLS method, which operates on the direction vectors and leads to sparse linear combinations of the original predictors given in terms of a surrogate vector (c). They define a two objective optimization problem where the weights are defined by the θ parameter, which controls the effect of the concavity of the objective function and the closeness of the original vector (w) and the surrogate direction vector (c)

$$\mathrm{Min}_{w,c} \; -\theta w'Mw + (1-\theta)(c-w)'M(c-w) + \lambda |c|_1 \tag{16}$$

subject to $w'w = 1$.

The additional term that appears in the optimization problem is given in terms of λ, the sparsity parameter which is a penalty that encourages sparsity on the direction vector. When θ=1, the first term is the original eigenvalue problem of PLS if $M=X'YY'X$, and of PC if $M=X'X$. When $M=X'X$, the problem becomes that of SCoTLASS when w=c and SPCA when θ=1/2 (see Zou and Hastie, 2005).

We include SPLS into the economic analysis. In particular, we explore its usefulness in the macroeconomic forecasting area, so we consider $M=X'YY'X$. Since only its static version is available in the literature, we also consider its extension to the dynamic case. In fact, we apply this methodology for the same alternative approaches proposed for PLS, and include dynamics in the sparse version of PLS (DSPLS).

5. EMPIRICAL APPLICATION

To check how the different procedures perform in terms of forecasting accuracy, we use the Stock and Watson database (2005) and an updated version of it. The target variable is the US logarithm of the Consumer Price Index, which is assumed to be integrated of order 2 (Stock and Watson, 2002b and Bai and Ng, 2008) and is defined as:

$$y_{t+h}^{h} = \frac{1200}{\mathrm{h}}(y_{t+h} - y_t) - 1200(y_t - y_{t-1}). \tag{17}$$

The forecasting model is estimated at each period as a function of its own lags $\phi(L)\ y_t$ and the estimated factors (F_t) and their lags. The parameters and factors are estimated with information up to time t (X_t and y_{t+h-1}). The number of lags of the predictors is chosen by the Bayesian Information Criterion (BIC). We consider several forecast horizons h=1, 6, 12 and 24 to check the performance of the different approaches in the short and medium run. The final forecasts are obtained as follow:

$$y_{t+h} = \mu + \phi'(L)y_t + \beta'(L)\hat{F}_t \tag{18}$$

It is important to state that instead of selecting some particular number of factors, derived from a particular criterion, we extract different number of factors from the data set and allow the final number of factors to be determined by the forecasting performance.

The original data set consists of 132 monthly United States (U.S.) macroeconomic time series that span the period from January 1960 through December 2003, for a total of T=528 observations.

The series are transformed to achieve stationarity by taking logs, first or second differences as necessary, as in Bai and Ng (2008) and Stock and Watson (2006).

For comparison purposes we employ seven forecast subsamples, as defined by Bai and Ng (2008), which can account for the temporal instability in the relation between the predictors and the variable to forecast. For factor estimation, the initial period of the dataset is always March 1960, whereas the final period is recursively expanded from February 1970 to February 1980 and February 1990 onwards until the end of each sample. The estimation and forecast samples are summarized in Table 1:

Table 1
Estimation and forecast subsamples

SS	Estimation subsample	Forecast subsample
M1	1960:03 to 1970:03-h	1970:03 to 1980:12
M2	1960:03 to 1980:03-h	1980:03 to 1990:12
M3	1960:03 to 1990:03-h	1990:03 to 2000:12
M4	1960:03 to 1970:03-h	1970:03 to 1990:12
M5	1960:03 to 1970:03-h	1970:03 to 2000:12
M6	1960:03 to 1980:03-h	1980:03 to 2000:12
M7	1960:03 to 1970:03-h	1970:03 to 2003:12

5.1 Forecast results

The predictive ability of the PC, TP, PLS and SPLS methods over a univariate benchmark is compared in tables 2 to 5 for the different forecast horizons considered. We use as benchmark an AR(4) for h=1. For the remaining forecast horizons we also regress y_{t+h} over y_t and three lags. As the measure for forecast comparison, we use the relative mean-squared forecast errors (RMSE) over the benchmark:

$$RMSE\ (method) = \frac{MSE\ (method)}{MSE\ (AR(4))}. \qquad (19)$$

An entry of less than one implies an improvement of the method upon the simple AR(4) forecast.

As regards PC regression, we try k=1 to 10 for the number of factors. Their lags, as well as the number of lags for the target variable, are selected by the BIC. We also borrow some of the forecasting results from Bai and Ng (2008). In particular, we consider the relative mean square forecast errors from the TP, in which the factors are estimated from a subset of the available data, using hard and soft threshold rules, and from the PC method where factors are estimated from the whole data set of predictors.

The PLS approach is implemented in the different versions considered in section 3 in order to take into account the properties of the data. With the aim of evaluating the SPLS forecast performance, we estimate the latent SPLS components, considering values for the sparsity parameter λ in the set {0.2, 0.4, 0.6 and 0.8}. The number of components considered is k=1and 2, although we have tried up to 5 components. Since the best forecasting results were obtained most of the time with just two components, we perform a more complete analysis for k=1 and 2.

Tables 2 to 5 show the forecasting results for h=1, 6, 12 and 24. They suggest some interesting observations of the competing methods. First, the results highlight that it is possible to make refinements to the factor forecasting methodology. We find efficiency gains over the widely used PC and over PLS by estimating sparse factors predictors by TP and SPLS. Second, SPLS is systematically the best procedure for the one month, six month and twelve month forecast horizons, as shown in column 8 of table 2 and column 7 of table 3 and 4, respectively. For these horizons, over 70% of the subsample periods yield the most precise forecast and for the rest of them its accuracy is similar to the best TP models. Third, in general, the improvements upon the benchmark are larger for longer forecast horizons. The better forecasting performance at longer horizons has also been found in the factor model literature (see, for instance, Matheson, 2006 and Caggiano et al, 2009).

Fourth, as regards the performance of PLS, it is important to note that when the dynamic relationship of the target variable is directly captured in the forecasting equation through the lagged values of the target variable (options a and c), the method provides better results, outperforming the benchmark and on a few occasions PC and TP too. Nevertheless, when the lags of the forecasting variable are incorporated as additional predictors in the dataset, (option b), the method performs even worse than the benchmark. The reason is that PLS gives weights to all the predictors, and then since the dimension of the cross section N is large, the weight given to y_t and its lags is weakened with respect to the options in which they are included directly in the forecasting equation. Notice, however that this is not necessarily the case with its sparse version, where this option performs very well in the short run (for h=1) in almost all samples, which can be seen from column 8 of table 2. In the particular case of h=1, the lags of the variable have a large predictive power with respect to other explanatory variables. The selection process seems to weight appropriately the relevant information for the prediction purpose, disregarding variables that have a negligible effect on the response and enough weight is given to y_t and its lags in order to capture the dynamic behavior of y_{t+h}.

Table 2
RMSE, h=1

Period	Bai and Ng (2008)		PLS			SPLS		
	PC (10)	Targeted Predictors	Option a (k=2)	Option b (k=1)	Option c (k=1)	Option a (k=2)	Option b (k=2)	Option c (k=1)
70.3-80.12	1.005	**0.944**	1.037	1.230	0.970	0.977	0.989	**0.955**
80.3-90.12	0.967	0.873	0.982	1.356	1.012	0.860	**0.825**	0.866
90.3-00.12	0.927	0.938	0.915	1.393	0.926	0.800	**0.797**	0.801
70.3-90.12	0.995	**0.938**	1.031	1.348	1.015	0.970	**0.950**	0.962
70.3-00.12	0.982	0.947	1.013	1.362	1.003	0.939	**0.922**	0.931
80.3-00.12	0.955	0.895	0.967	1.376	0.993	0.839	**0.813**	0.843
70.3-03.12	0.974	0.937	1.000	1.374	0.994	0.935	**0.920**	0.931

Source: Bai and Ng (2008) and authors' calculations. The table shows the ratio of MSE of PC, TP, PLS and SPLS over the benchmark model for h=1. PLS and SPLS results are shown only for the number of components k which yields the best forecasting results.

Table 3
RMSE, h=6

Period	Bai and Ng (2008)		PLS			SPLS		
	PC (10)	Targeted Predictors	Option a (k=2)	Option b (k=1)	Option c (k=1)	Option a (k=2)	Option b (k=2)	Option c (k=1)
70.3-80.12	0.712	0.665	0.609	1.437	0.561	**0.497**	0.545	0.558
80.3-90.12	0.654	**0.571**	0.634	2.147	0.638	**0.601**	0.626	0.606
90.3-00.12	0.660	0.651	0.727	2.413	0.735	**0.547***	0.765	0.680
70.3-90.12	0.675	0.608	0.656	1.899	0.631	**0.583**	0.638	0.621
70.3-00.12	0.671	0.610	0.650	1.935	0.627	**0.579**	0.638	0.615
80.3-00.12	0.652	**0.582**	0.628	2.159	0.631	**0.592**	0.642	0.598
70.3-03.12	0.670	0.609	0.645	1.977	0.632	**0.586**	0.646	0.614

Source: Bai and Ng (2008) and authors´ calculations. The table shows the ratio of MSE of PC, TP, PLS and SPLS over the benchmark model for h=6. An asterisk (*) means k=1.

Fifth, the SPLS, option a, produces the best results for h=6 and h=12, which indicates that for these horizons some (but not all) the predictor variables, as distinct to the lags of the target, contain relevant information about the variable to forecast. Accordingly, the number of variables chosen by the SPLS options is larger for longer horizons.

Sixth, for the larger forecast horizon considered (h=24), the best performance SPLS models includes a large number of components, k=3 to k=5. According to the estimated RMSE reported in table 5 (column 7), SPLS (option a) outperforms PC in all the subsamples and do slightly better than TP in two of them, while for the remaining subsamples its prediction accuracy is similar to the best TP models.

Table 4
RMSE, h=12

Period	Bai and Ng (2008)		PLS			SPLS		
	PC (10)	Targeted Predictors	Option a (k=2)	Option b (k=1)	Option c (k=1)	Option a (k=2)	Option b (k=2)	Option c (k=2)
70.3-80.12	0.631	0.580	0.636	0.926	0.574	**0.537***	0.624	0.557
80.3-90.12	0.575	0.560	0.807	1.277	0.649	**0.536**	0.553	0.551
90.3-00.12	0.723	0.616	0.904	1.059	0.882	0.630	**0.591**	0.744
70.3-90.12	0.603	0.573	0.722	1.104	0.613	**0.556**	0.601	0.573
70.3-00.12	0.611	**0.573**	0.730	1.093	0.629	**0.578**	0.621	0.585
80.3-00.12	0.594	**0.568**	0.806	1.216	0.670	**0.568**	0.584	0.576
70.3-03.12	0.609	**0.570**	0.716	1.085	0.625	0.573	0.612	0.612

Source: Bai and Ng (2008) and authors´ calculations. The table shows the ratio of MSE of PC, TP, PLS and SPLS over the benchmark model for h=12. PLS and SPLS results are shown only for the number of components k that yields the best forecasting results. An asterisk (*) means k=1.

Table 5
RMSE, h=24

Period	Bai and Ng (2008)		PLS			SPLS		
	PC (10)	Targeted Predictors	Option a (k=5)	Option b (k=1)	Option c (k=1)	Option a (k=3)	Option b (k=2)	Option c (k=1)
70.3-80.12	0.532	0.486	0.491	1.076	0.968	**0.436****	1.030	0.958
80.3-90.12	0.506	**0.431**	0.860	1.836	0.989	**0.479**	1.005	0.991
90.3-00.12	0.546	**0.447**	0.971*	1.146	0.965	**0.696**	0.972	0.967
70.3-90.12	0.522	0.467	0.675	1.435	0.978	**0.464***	1.019	0.976
70.3-00.12	0.523	**0.464**	0.705	1.404	0.977	**0.488***	1.013	0.976
80.3-00.12	0.512	**0.440**	0.887	1.570	1.003	**0.507**	1.035	1.004
70.3-03.12	0.523	**0.464**	0.708	1.405	0.967	**0.493***	1.059	0.967

Source: Bai and Ng (2008) and authors´ calculations. The table shows the ratio of MSE of PC, TP, PLS and SPLS over the benchmark model for h=24. PLS and SPLS results are shown only for the number of components k that yields the best forecasting results. An asterisk (*) means k=4 and ** denotes k=5.

5.2 The variables chosen

The fifth column of table 6 shows that the average number of variables selected for the SPLS in the case of h=1 is small, very close to the one estimated by Bai and Ng (2008) for the number of best predictor variables (k*(LARS)).

The number of variables chosen period by period over all samples oscillates in the range between 1 and 104, due to the instability of the forecast period 1970 to 1980 that constitutes the first sample. However, when this sample is excluded, the interval is significantly reduced to 1 to 15 variables selected for the rest of the samples considered. For the best performing option in each model, the average number of chosen variables decreases from a range between 1 to 85, when considering all the forecasting subsamples, to a range between 1 to 5, as shown in columns 4 through 6 of table 6. The outcomes imply a high degree of sparsity, λ=0.8 in all the forecasting samples except the first, for h=1.

For options a and c, when the lags of the target variable are excluded from the set of predictors and included explicitly in the forecasting equations, the most frequently selected variables are related to interest rates: Treasury Bonds (TB), Fed Funds, employment and the services component of the CPI.

In what follows we focus solely on the selection of variables for the best forecasting results. When the lags of the target variable are included in the expanded set of predictors X_e, as happens in option b, the first lag of the CPI has a sizable contribution to forecasting CPI in the short term. For example, in the third forecasting sample that includes the period from 1990 to 2000, a stable decade of relatively mild inflation, the method selects only the first lag and the services price component as relevant variables. This phenomenon has been associated to improved monetary policy making, a result of smaller and more infrequent shocks hitting the economy and a structural break in the relationship between the inflation and the common factors, which constitute the most frequently explanations for The Great Moderation (Bernanke, 2004; Summers, 2005; Kim et al., 2004; among others). For the rest of samples, the predictors selected for the shortest term horizon (h=1) can be systematically grouped into three prevailing categories: monetary, unemployment and price components. The variables that dominate the list are CPI services, monetary base and interest rates: Fed Funds, 5 years and 1 year Treasury Bills.

Table 6
Average number of selected variables. h=1

Sample	Targeted Predictors* (TPC)	PC	SPLS method			k*(LARS)
			Option A (k=2)	Option B (k=2)	Option C (k=1)	
70.1-80.1	**32.174**	132	45.438	48.677	85.323	7.694
80.1-90.1	30.000	132	2.153	**2.177**	3.008	5.661
90.1-00.1	73.884	132	2.000	**2.000**	1.000	7.273
70.1-90.1	**30.000**	132	3.628	**4.128**	4.640	6.689
70.1-00.1	30.000	132	3.100	**3.438**	3.459	6.886
80.1-00.1	68.154	132	3.208	**3.216**	4.889	6.469
70.1-03.9	30.000	132	3.005	**3.314**	3.247	7.038

Source: Bai and Ng (2008) and authors´ calculations. The table shows the average number of selected variables with the different methods applied to form the linear combinations that constitute the unobserved “factors”.

As we have mentioned before, the number of variables chosen by the SPLS options grows for longer horizons, due to growing uncertainties and the necessity to account for other possible sources of variability to explain the behavior of the target variable; this is evident by comparing the average number of selected variables for h=1, h=6 and h=12, reported in the fifth column of table 6 and the fourth columns of tables 7 and 8, respectively.

Table 7
Average number of selected variables. h=6

Sample	Targeted Predictors* (TPC)	PC	SPLS method			k*(LARS)
			Option A (k=2)	Option B (k=2)	Option C (k=1)	
70.1-80.1	30.000	132	**22.777**	21.354	79.823	7.694
80.1-90.1	**30.000**	132	34.277	4.969	12.131	5.661
90.1-00.1	30.000	132	**11.900**	11.900	104.938	7.273
70.1-90.1	30.000	132	**28.500**	28.300	79.384	6.689
70.1-00.1	30.000	132	**30.978**	31.289	78.581	6.886
80.1-00.1	**30.000**	132	35.172	4.024	45.192	6.469
70.1-03.9	30.000	132	**30.748**	30.820	78.465	7.038

Source: Bai and Ng (2008) and authors´ calculations. The table shows the average number of selected variables by each method for h=6.

In the case of h=6, the number of variables oscillates between 4 and 104 and the average fluctuated between 12 and 35. For this horizon, the degree of sparsity is lower than the observed for h=1 (λ=0.6, for all but the third sample where λ= 0.4). The selected variables are different from the ones selected for h=1. The money supply (real M2) is a variable selected in all samples and at each t in each sample. The unemployment is a predominant group influencing the behavior of the target followed by price components. There are some variables related to the demand, such as the consumption variable and the Purchasing Managers' Index (PMI).

For h=12, the chosen variables range from 6 to 105 and the sparsity parameter changes with the samples from 0.2 to 0.6 and 0.8. The relative stability/instability of the target variable in the forecast period influences this result significantly. The money supply (real M2) appears to have strong predictive power for inflation and is the most frequently selected variable, such as for h=6, although the unemployment and production variables are the dominant for this horizon.

Table 8
Average number of selected variables. h=12

Sample	Targeted Predictors* (TPC)	PC	SPLS method			k*(LARS)
			Option A (k=2)	Option B (k=2)	Option C (k=2)	
70.1-80.1	30.000	132	**59.262**	9.746	22.623	14.992
80.1-90.1	5.000	132	**35.825**	36.385	40.677	10.727
90.1-00.1	30.000	132	4.177	**4.615**	69.377	11.769
70.1-90.1	30.000	132	**30.816**	10.308	36.484	12.867
70.1-00.1	**30.000**	132	**32.454**	8.438	36.095	12.504
80.1-00.1	**30.000**	132	**7.792**	36.616	38.088	11.249
70.1-03.9	**30.000**	132	**32.703**	32.624	35.891	12.769

Source: Bai and Ng (2008) and authors´ calculations. The table shows the average number of selected variables by each method for h=12.

Table 9 shows the average number of selected variables. Taking into account the best performing option in each model for h=24, the average number ranges from 7 to 84. The degree of sparsity for options a and b is 0.8 for all the samples, while for option c it oscillates between 0.2 and 0.6. The money supply (real M2), as in the case of h=6, was selected in all samples and in all forecasting periods. The economic activity variables were repeatedly selected and accounted for 60% of the top ten for this horizon.

Table 9
Average number of selected variables. h=24

Sample	Targeted Predictors* (TPC)	PC	SPLS method			k*(LARS)
			Option A (k=4)	Option B (k=2)	Option C (k=2)	
70.1-80.1	30.000	132	**16.362**	7.662	21.862	15.355
80.1-90.1	**30.000**	132	**13.385**	8.908	83.615	14.917
90.1-00.1	**30.000**	132	**7.964**	2.130	81.623	15.025
70.1-90.1	30.000	132	**17.208**	8.244	26.100	15.158
70.1-00.1	**30.000**	132	**16.349**	7.373	26.757	15.119
80.1-00.1	**30.000**	132	**10.504**	6.735	59.365	14.979
70.1-03.9	**30.000**	132	**16.178**	7.116	26.807	15.119

Source: Bai and Ng (2008) and authors´ calculations. The table shows the average number of selected variables by each method for h=24.

5.3 Other variables

To test the empirical validity of the sparse factor models, we apply the proposed procedures to the series: Industrial Production (IP) and Total Employment (EMT). Like Bai and Ng (2008), we report the results only for h=12 and assume that the log level of the two series are differenced stationary. The target variables are defined as follows:

$$y^h_{t+h} = \frac{1200}{\mathrm{h}}(y_{t+h} - y_t) \text{ and } z_t = 1200(y_t - y_{t-1}). \tag{20}$$

The results are reported in Tables 10 and 11. First, we find that for both series SPLS methods outperform the standard PC most of the time - columns 2 and 8 of table 10 show the corresponding comparisons for IP, and columns 2 and 7 of table 11 report those for EMT. Second, when we compare the SPLS forecast performance to TP, we find forecast improvements in 50% of the cases. This result supports the findings for CPI about the possibility of obtaining gains in terms of statistical accuracy by the employment of sparse factor models.

Table 10
RMSE, IP, h=12

Period	Bai and Ng (2008)		PLS			SPLS		
	PC (10)	Targeted Predictors	Option a (k=1)	Option b (k=1)	Option c (k=1)	Option a (k=1)	Option b (k=2)	Option c (k=2)
70.3-80.12	0.247	**0.197**	0.227*	1.196*	0.621	**0.232***	0.634	0.612
80.3-90.12	0.846	0.692	1.516	2.051	1.290	0.830	**0.553**	1.188
90.3-00.12	1.055	1.327	2.325*	2.552	1.832	1.116*	**0.591**	1.798
70.3-90.12	0.442	**0.359**	0.728	1.721	0.566	**0.498**	0.606	0.513
70.3-00.12	0.497	**0.456**	0.870	1.781	0.626	0.594	0.603	**0.575**
80.3-00.12	0.898	0.866	1.672	1.814	1.239	1.059	**0.768**	1.129
70.3-03.12	0.551	**0.513**	0.988	1.843	0.696	0.667	**0.643**	0.645

Source: Bai and Ng (2008) and authors´ calculations. The table shows the ratio of MSE of PC, TP, PLS and SPLS over the benchmark model for h=12. PLS and SPLS results are shown only for the number of components k that yields the best forecasting results. An asterisk (*) means k=2

We apply the methodology for k=1 and k= 2 components, given that in most cases the best forecasting results were obtained with a small number of components. The degree of sparsity is high in all samples and models (λ=0.8) with the exception of the first one. As mentioned before, the instability of this forecast period seems to require more variables to account for the variance of the target.

Table 11
RMSE, EMT, h=12

Period	Bai and Ng (2008)		PLS			SPLS		
	PC (10)	Targeted Predictors	Option a (k=2)	Option b (k=1)	Option c (k=1)	Option a (k=2)	Option b (k=2)	Option c (k=2)
70.3-80.12	0.524	**0.383**	0.586*	1.197	0.574	**0.537***	0.634	0.604*
80.3-90.12	0.644	0.591	0.807	2.051	0.649	**0.536**	0.553	0.551
90.3-00.12	0.947	0.965	0.904	2.552	0.882	0.630	**0.591**	0.744
70.3-90.12	0.569	**0.459**	0.706*	1.721	0.613	**0.556**	0.606	0.573
70.3-00.12	0.616	**0.545**	0.730	1.781	0.629	**0.578**	0.603	0.585
80.3-00.12	0.730	0.744	0.835	1.814	0.795	**0.638**	0.768	0.730
70.3-03.12	0.696	0.609	0.718	1.843	0.628	**0.579**	0.643	0.580

Source: Bai and Ng (2008) and authors´ calculations. The table shows the ratio of MSE of PC, TP, PLS and SPLS over the benchmark model for h=12. PLS and SPLS results are shown only for the number of components k that yields the best forecasting results. An asterisk (*) means k=1.

5.4 Updated dataset

We perform an update of the Stock and Watson (2005) dataset. The updated base contains 112 monthly macroeconomic time series, and extends the time series of the original base through December 2010 for a total of T= 610 observations. For comparison purposes, we divide the updated dataset into the three subsamples shown in table 12:

Table 12
Estimation and forecast subsamples

SS	Estimation subsample	Forecast subsample
M1	1960:03 to 1970:03-h	1970:03 to 2003:12
M2	1960:03 to 2000:02-h	2000:02 to 2010:12
M3	1960:03 to 1970:03-h	1970:03 to 2010:12

The initial period for factor estimation is always March 1960, as in the previous application, whereas the final periods are recursively expanded from February 1970 and January 2000 until the end of the setup. As a reference, the first updated subsample coincides with the last one considered for the original database.

We implement the three different versions of the PLS and SPLS approaches proposed in section 3 and the standard Principal Components (PC (10)). Table 13 summarizes the forecasting results for h=1, 6, 12 and 24. The main findings regarding this update are the following: SPLS seems the best forecasting procedure for h=1, 6, 12 and 24; in 75% of the samples it produces the most accurate forecast, as columns 7 and 8 confirm. Second, the improvements over the benchmark are larger for sample 2, which means that SPLS provides a better forecast for the 2000's decade. Third, SPLS gives better results for options a and c, where the dynamics of the target are taken into account in the forecasting equation rather than in the selection method.

Table 14 shows the average number of variables selected. As in the previous cases, when the forecasting sample includes the 70s, the number of variables is larger, while it is reduced by around 10 for the XXI century.

Table 13
RMSE, CPI

Period	PC (10)	PLS			SPLS		
		Option a	Option b	Option c	Option a	Option b	Option c
h=1							
70.3-03.12	0.974	0.962	1.195	**0.943**	0.976	1.074	**0.943**
00.2-10.12	0.809	0.838	1.100	0.838	0.814	1.075	**0.779**
70.3-10.12	**0.905**	0.977	1.085	0.969	0.960	0.977	**0.950**
h=6							
70.3-03.12	0.670	0.618	1.135	0.621	**0.604**	0.930	0.622
00.2-10.12	0.969	0.719	1.063	0.723	**0.621**	1.037	0.724
70.3-10.12	0.791	0.721	0.931	0.744	**0.621**	0.728	0.745
h=12							
70.3-03.12	0.609	0.592	1.093	0.591	0.581	0.892	**0.577**
00.2-10.12	0.709	0.622	1.159	0.605	0.616	1.025	**0.573**
70.3-10.12	0.605	0.616	0.979	0.614	**0.604**	0.653	0.612
h=24							
70.3-03.12	0.696	0.572	1.064	0.550	**0.484**	0.822	0.541
00.2-10.12	0.902	0.640	1.295	0.620	0.629	1.137	**0.617**
70.3-10.12	**0.481**	0.594	1.079	0.581	**0.535**	0.692	0.574

Source: Authors´ calculations and Bai and Ng (2008) for PC(10) in the first subsample. The table shows the ratio of MSE of PC, PLS and SPLS over the benchmark model for h=1, 6, 12 and 2. PLS and SPLS results are shown only for the number of components k that yields the best forecasting results.

Table 14
Average number of selected variables for CPI

Horizon / Sample	h=1		h=6		h=12		h=24	
	PC	SPLS	PC	SPLS	PC	SPLS	PC	SPLS
70.3-03.12	132	80.430	132	27.936	132	30.195	132	23.543
00.2-10.12	112	11.212	112	10.519	112	11.121	112	10.756
70.3-10.12	112	9.178	112	26.004	112	59.106	112	23.871

Source: Authors´ calculations. The table shows the average number of variables selected by each method for h=1, 6, 12 and 24.

6. CONCLUSIONS

The empirical results are encouraging, suggesting that there is some room for refinement the factor forecasting methodology. The dynamic SPLS methodology introduced in this paper shows a good prediction performance, improving the forecast efficiency of the alternative widely used factor methods in macroeconomic forecasting. Our findings confirm that the choice of a useful or informative subset of predictors, to extract the latent variables to forecast a specific target variable is relevant for improving the performance of the factor forecasting methods. More variables (more information) do not necessarily yield better forecasting results.

Among the different possibilities analyzed to apply PLS and SPLS to time series data, it seems that applying directly the PLS techniques between the target variable and the predictors yields the better forecasting results. Enlarging the data set of predictors, by including the lags of the target variable in it, does not seem to be a good alternative for PLS when applied to time series data, although this is not necessarily the case when the sparse version is applied. The PLS method gives weight to all the forecasting predictors, so the dependence between the target variable and its past can be obscured if there are too many predictors. On the contrary, including the lags of the target variable explicitly on the forecasting equation seems to be the best way of capturing the dynamic behavior of the target.

In the short run (h=1), the dynamic SPLS approaches perform very well. When the dynamic relationship is integrated through the inclusion of the lags of the target as additional predictors in the original dataset, the selection process seems to weight the relevant information for forecasting purposes appropriately. In particular, the presence of variables that have a negligible effect on the response do not lessen the participation of y_t and its lags. For the updated dataset, the isolation of the AR(p) process effects, before PLS estimation, shows a good performance for h=1 and also for some subsamples in larger horizons.

The variable selection performed by the SPLS model shows differences between the periods of high and low uncertainty in the economic environment and between the forecasting horizons and thus evidences the relevance of increasing the flexibility in the factor forecasting methodology. The proposed SPLS method has more flexibility than the traditional benchmarks; it allows choosing suitable predictors period by period to forecast a target and monitoring the variables that go in/out the model, so it can also be used as an exploratory tool.

Additionally, the variables chosen by the SPLS model in the CPI case have an economic foundation. The variables chosen to forecast inflation are mainly monetary variables: interest rate and monetary aggregates (real M2 or monetary base), price components and real activity variables: unemployment, housing starts, industrial sector activity indicators. There are some variables associated to the demand side such as the consumption variable, consumption credit and some components of PMI and of the National Association of Purchasing Managers' index (NAMP). A greater interpretability of the results is an additional gain of the proposed methodology.

REFERENCES

Bai, J. and S. Ng (2002). “Determine the Number of Factors in Approximate Factor Models”. Econometrica 70: 191-221.

_________________ (2008). “Forecasting Economic Time Series Using Targeted Predictors”. *Journal of Econometrics*, 146:304-317.

Bernanke, B. (2004). “The Great Moderation’, Remarks by B. Bernanke Member of the Board of Governors of the U.S. Federal Reserve System, at the meetings of the Eastern Economic Association, Washington, D.C. and BIS Review 12/2004.

Boivin, J. and S. Ng (2005). “Understanding and Comparing Factor-Based Forecasts”. *International Journal of Central Banking*, Vol. 1(3).

___________________(2006). “Are more Data Always Better for Factor Analysis?” *Journal of Econometrics*, 132: 169-194.

Caggiano, G., Kapetanios G., and V. Labhard (2009). “Are More Data Always Better for Factor Analysis? Results for the Euroarea, the Six largest Euroarea Countries and UK”, Working Paper Series No.1051. European Central Bank.

Chun H. and S. Keles (2010). “Sparse Partial Squares Regression for Simultaneous Dimension Reduction and Variable Selection”, *Journal of the Royal Statistical Society*. Series B, Statistical Methodology, 72(1): 3-25.

De Mol, C., Giannone, D. and L. Riechlin (2008). “Forecasting Using a Large Number of Predictors. Is Bayesian Regression a Valid Alternative to Principal Components?”. *Journal of Econometrics*, Vol. 146 (2):318-328.

Doz, C., Giannone, D. and L. Riechlin (2012). “A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor”, *Review of Economics and Statistics*, forthcoming.

Eickmeier, S. and T. Ng (2011). “Forecasting National Activity using lots of International Predictors: An Application to New Zealand”. *International Journal of Forecasting*, Vol. 27(2):496-511.

Engle R. and M. Watson (1981). “A One Factor Multivariate Time Series Model of Metropolitan Wage Rates”, *Journal of the American Statistical Association*, 76:774-781.

Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2000). “The Generalized Dynamic Factor Model: Identification and Estimation”. *The Review of Economics and Statistics* (82) 4: 540-554.

____ (2005). "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting." *Journal of the American Statistical Association* 100 (471): 830–40.

Geweke, J. (1977). "The Dynamic Factor Analysis of Economic Time Series", in *Latent Variables in Socio Economic Models,* ed. by D.J. Aigner and A.S. Goldberger, Amsterdam: North Holland.

Geweke J. and K. Singleton (1981). "Latent Variable Models for Time Series: A Frequency Domain Approach with an Application to the Permanent Income Hypothesis". *Journal of Econometrics*, Vol. 17(3): 287-304.

Groen, J. and G. Kapetanios (2008). "Revisiting useful approaches to data-rich macroeconomic forecasting", Federal Reserve Bank of New York Staff Report 327.

Hastie T., Tibshirani, R. and J. Friedman (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics, 2nd. edition.

Inoue, A., and L. Kilian (2008). "How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," *Journal of the American Statistical Association*, Vol. 103: 511 – 522.

Jungbacker, B. and S.J. Koopman (2008). "Likelihood-based Analysis for Dynamic Factor Models". Discussion Paper, Tinbergen Institute. TI 2008-0007/4.

Kim, C.J., Nelson, C. and J. Piger (2004). "The Less Volatile U.S. Economy: A Bayesian Investigation on Timing, Breadth, and Potential Explanations", Journal of Business and Economic Statistics, Vol. 22: 80-93.

Mathenson (2006). "Factor Model Forecast for New Zealand". *International Journal of Central Banking*, Vol. 2 (2).

Stock, J. and M. Watson (2002a). "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association*, 97 (460): 1167–79.

____________________ (2002b). "Macroeconomic Forecasting Using Diffusion Indexes". *Journal of Business & Economic Statistics*, 20, 147-163.

____________________ (2005). "Implications of Dynamic Factor Models for VAR Analysis". NBER Working Paper No. 11467.

____________________ (2006). "Forecasting with Many Predictors." In The *Handbook of Economic Forecasting*, Vol. 1, ed. G. Elliott, C. Granger and A. Timmermann. Elsevier Science.

____________________ (2007a). "Forecasting In Dynamic Factor Models Subject To Structural Instability". Prepared for the Conference in Honor of David Hendry, August 23-25, 2007, Oxford.

____________________ (2007b). "Why U.S. Inflation Become Harder to Forecast". *Journal of Money, Credit and Banking*, 39: 13-23

____________________ (2010). Dynamic Factor Models. *Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

Summers, P. (2005). “What Causes the Great Moderation? Some Cross-Country Evidence”, *Economic Review*, Federal Reserve Bank of Kansas City. 3rd quarter.

Wang, C. (2008). “Dimension reduction techniques for forecasting: An empirical comparison”. Mimeo.

Watson, M. (2000). “Macroeconomic Forecasting Using Many Predictors”, in M. Dewatripont, L. Hansen and S. Turnovsky (eds), Advances in Economics and Econometrics, Theory and Applications, Eight World Congress of the Econometric Society, Vol. III: 87-115.

Wold, H. (1966). “Estimation of Principal Components and Related Models by Iterative Least Squares”. In P.R. Krishnaiaah (Ed.), *Multivariate Analysis* (pp. 391-420). New York: Academic Press.

Zou, H., T. Hastie (2005). “Regularization and variable selection via the elastic net”, *Journal of Royal Statistical Society*, Series B, 67(2): 301-320.

Appendix A. Data definitions and transformations (Stock and Watson 2005)

Short name	Transformation	Mnemonic	Description
PI	Δln	DLPI	Personal income (AR, bil. chain 2000 $)
PI less transfers	Δln	DLPILTRANSFERS	Personal income less transfer payments (AR, bil. chain 2000 $)
Consumption	Δln	DLCONS	Real Consumption (AC) A0m224/gmdc
M&T sales	Δln	DLMTSALES	Manufacturing and trade sales (mil. Chain 1996 $)
Retail sales	Δln	DLRETAILSALES	Sales of retail stores (mil. Chain 2000 $)
IP: total	Δln	DLIPTOTAL	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX
IP: products	Δln	DLIPPRODUCTS	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL
IP: final prod	Δln	DLIPFINALPROD	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS
IP: cons gds	Δln	DLIPCONSGDS	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS
IP: cons dble	Δln	DLIPCONSDBLE	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS
IP: cons nondble	Δln	DLIPCONSNONDBLE	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS
IP: bus eqpt	Δln	DLIPBUSEQPT	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT
IP: materials	Δln	DLIPMATLS	INDUSTRIAL PRODUCTION INDEX - MATERIALS
IP: dble matls	Δln	DLIPDBLEMATLS	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS
IP: nondble matls	Δln	DLIPNONDBLEMATLS	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS
IP: mfg	Δln	DLIPMFG	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)
IP: res util	Δln	DLIPRESUTIL	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES
IP: fuels	Δln	DLIPFUELS	INDUSTRIAL PRODUCTION INDEX - FUELS
NAPM prodn	lv	NAPMPRODN	NAPM PRODUCTION INDEX (PERCENT)
Cap util	Δlv	DCAPUTIL	Capacity Utilization (Mfg)
Help wanted indx	Δlv	DHELPWANTDIND	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)
Help wanted/emp	Δlv	DHELPWANTEMP	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF
Emp CPS total	Δln	DLEMPCPSTOTAL	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)
Emp CPS nonag	Δln	DLEMPCPSNONAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)
U: all	Δlv	DUNEMPALL	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)
U: mean duration	Δlv	DUNMEANDUR	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)
U< 5 wks	Δln	DLUNL5WKS	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)
U 5-14 wks	Δln	DLUN514WKS	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)
U 15+ wks	Δln	DLUN15MWKS	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)
U 15-26 wks	Δln	DLUN1526WKS	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)
U 27+ wks	Δln	DLUN27MWKS	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS,SA)
UI claims	Δln	DLUICLAIMS	Average weekly initial claims, unemploy. insurance (thous.)
Emp: total	Δln	DLEMPTOTAL	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE
Emp: gds prod	Δln	DLEMPGDSPROD	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING
Emp: mining	Δln	DLEMPMINING	EMPLOYEES ON NONFARM PAYROLLS - MINING
Emp: const	Δln	DLEMPCONST	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION
Emp: mfg	Δln	DLEMPMFG	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING
Emp: dble gds	Δln	DLEMPDBLEGDS	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS
Emp: nondbles	Δln	DLEMPNONDBLES	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS
Emp: services	Δln	DLEMPSERV	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING
Emp: TTU	Δln	DLEMPTTU	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES
Emp: wholesale	Δln	DLEMPWHSALE	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE
Emp: retail	Δln	DLEMPRETAIL	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE
Emp: FIRE	Δln	DLEMPFIRE	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES
Emp: Govt	Δln	DLEMPGOV	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT
Emp-hrs nonag	Δln	DLEMPHNONAG	Employee hours in nonag. establishments (AR, bil. hours)
Avg hrs	lv	AVGHRS	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR
Overtime: mfg	Δlv	DOVERTMFG	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR
Avg hrs: mfg	lv	AVGHMFG	Average weekly hours, mfg. (hours)
NAPM empl	lv	NAPMEMP	NAPM EMPLOYMENT INDEX (PERCENT)
Starts: nonfarm	ln	LSTSNONFARM	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA
Starts: NE	ln	LSTSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.
Starts: MW	ln	LSTSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.
Starts: South	ln	LSTSSOUTH	HOUSING STARTS:SOUTH (THOUS.U.)S.A.
Starts: West	ln	LSTSWEST	HOUSING STARTS:WEST (THOUS.U.)S.A.
BP: total	ln	LBPTOTAL	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)
BP: NE	ln	LBPNE	HOUSES AUTHORIZED BY BUILD. PERMITS:NORTHEAST(THOU.U.)S.A
BP: MW	ln	LBPMW	HOUSES AUTHORIZED BY BUILD. PERMITS:MIDWEST(THOU.U.)S.A.
BP: South	ln	LBPSOUTH	HOUSES AUTHORIZED BY BUILD. PERMITS:SOUTH(THOU.U.)S.A.
BP: West	ln	LBPWEST	HOUSES AUTHORIZED BY BUILD. PERMITS:WEST(THOU.U.)S.A.
PMI	lv	PMI	PURCHASING MANAGERS' INDEX (SA)
NAPM new ordrs	lv	NAPMNWORD	NAPM NEW ORDERS INDEX (PERCENT)
NAPM vendor del	lv	NAPMVDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)
NAPM Invent	lv	NAPMINVT	NAPM INVENTORIES INDEX (PERCENT)

Short name	Transformation	Mnemonic	Description
Orders: cons gds	Δln	DLORDRCONGDS	Mfrs' new orders, consumer goods and materials (bil. chain 1982 $)
Orders: dble gds	Δln	DLORDRDBLGDS	Mfrs' new orders, durable goods industries (bil. chain 2000 $)
Orders: cap gds	Δln	DLORDRCAPGDS	Mfrs' new orders, nondefense capital goods (mil. chain 1982 $)
Unf orders: dble	Δln	DLUNORDDBLE	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 $)
M&T invent	Δln	DLMTINVENT	Manufacturing and trade inventories (bil. chain 2000 $)
M&T invent/sales	Δlv	DMTINVTSAL	Ratio, mfg. and trade inventories to sales (based on chain 2000 $)
M1	Δ2ln	DL2M1	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)
M2	Δ2ln	D2LM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,
M3	Δ2ln	DL2M3	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S&INST ONLY MMMFS)(BIL$,SA)
M2(real)	Δln	DLM2REAL	MONEY SUPPLY - M2 IN 1996 DOLLARS (BCI)
MB	Δ2ln	DL2MB	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)
Reserves tot	Δ2ln	DL2RESERVTOT	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)
Reserves nonbor	Δ2ln	DL2RESERVNONBOR	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL$,SA)
C&I loans	Δ2ln	DL2CILOANS	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1996 DOLLARS (BCI)
D C&I loans	lv	DELTACILOANS	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)
Cons credit	Δ2ln	DL2CONSCREDIT	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)
Inst credit/ PI	Δlv	DINSTCREDPI	Ratio, consumer installment credit to personal income (pct.)
S&P 500	Δln	DLSP500	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)
S&P: indust	Δln	DLSPINDUST	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)
S&P div yield	Δlv	DSPDIVYIELD	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)
S&P PE ratio	Δln	DLSPPERATIO	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)
Fed Funds	Δlv	DFEDFUNDS	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)
Comm paper	Δlv	DCOMPAPER	Cmmercial Paper Rate (AC)
3 mo T-bill	Δlv	DTBILL3M	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)
6 mo T-bill	Δlv	DTBILL6M	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)
1 yr T-bond	Δlv	DTBOND1Y	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)
5 yr T-bond	Δlv	DTBOND5Y	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)
10 yr T-bond	Δlv	DTBOND10Y	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)
Aaa bond	Δlv	DAAABOND	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)
Baa bond	Δlv	DBAABOND	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)
CP-FF spread	lv	CPFFSPREAD	cp90-fyff
3 mo-FF spread	lv	FFSPREAD3M	fygm3-fyff
6 mo-FF spread	lv	FFSPREAD6M	fygm6-fyff
1 yr-FF spread	lv	FFSPREAD1Y	fygt1-fyff
5 yr-FF spread	lv	FFSPREAD5Y	fygt5-fyff
10 yr-FF spread	lv	FFSPREAD10Y	fygt10-fyff
Aaa-FF spread	lv	AAAFFSPREAD	fyaaac-fyff
Baa-FF spread	lv	BAAFFSPREAD	fybaac-fyff
Ex rate: avg	Δln	DLEXRATEAVG	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)
Ex rate: Switz	Δln	DLEXRATESWITZ	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)
Ex rate: Japan	Δln	DLEXRATEJAPAN	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)
Ex rate: UK	Δln	DLEXRATEUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)
Ex rate: Canada	Δln	DLEXRATECANADA	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)
PPI: fin gds	Δ2ln	DL2PPIFINGDS	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)
PPI: cons gds	Δ2ln	DL2PPICONSGDS	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)
PPI: int mat'ls	Δ2ln	DL2PPIINTMATLS	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)
PPI: crude mat'ls	Δ2ln	DL2PPICRUDEMAT	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)
Spot market price	Δ2ln	DL2SPOTMKPRICE	SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100)
Sens mat'ls price	Δ2ln	DL2SENSIMATPRICES	INDEX OF SENSITIVE MATERIALS PRICES (1990=100)(BCI-99A)
NAPM com price	lv	NAPMCOMPRICE	NAPM COMMODITY PRICES INDEX (PERCENT)
CPI-U: all	Δ2ln	DL2PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)
CPI-U: apparel	Δ2ln	DL2CPIUAPPAREL	CPI-U: APPAREL & UPKEEP (82-84=100,SA)
CPI-U: transp	Δ2ln	DL2CPIUTRANSP	CPI-U: TRANSPORTATION (82-84=100,SA)
CPI-U: medical	Δ2ln	DL2CPIUMEDICAL	CPI-U: MEDICAL CARE (82-84=100,SA)
CPI-U: comm.	Δ2ln	DL2CPIUCOMM	CPI-U: COMMODITIES (82-84=100,SA)
CPI-U: dbles	Δ2ln	DL2CPIUDBLES	CPI-U: DURABLES (82-84=100,SA)
CPI-U: services	Δ2ln	DL2CPIUSERVICES	CPI-U: SERVICES (82-84=100,SA)
CPI-U: ex food	Δ2ln	DL2CPIUEXFOOD	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)
CPI-U: ex shelter	Δ2ln	DL2CPIUEXSHEL	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)
CPI-U: ex med	Δ2ln	DL2CPIUXMED	CPI-U: ALL ITEMS LESS MEDICAL CARE (82-84=100,SA)
PCE defl	Δ2ln	DL2PCEDEFL	PCE,IMPL PR DEFL:PCE (1987=100)
PCE defl: dbles	Δ2ln	DL2PCEDEFLDUR	PCE,IMPL PR DEFL:PCE; DURABLES (1987=100)
PCE defl: nondble	Δ2ln	DL2PCEDEFNONDUR	PCE,IMPL PR DEFL:PCE; NONDURABLES (1996=100)
PCE defl: service	Δ2ln	DL2PCEDEFSERVICE	PCE,IMPL PR DEFL:PCE; SERVICES (1987=100)
AHE: goods	Δ2ln	DL2AHEGOODS	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO
AHE: const	Δ2ln	DL2AHECONST	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO
AHE: mfg	Δ2ln	DL2AHEMFG	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NO
Consumer expect	Δlv	DCONSEXP	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)

In the transformation column, ln denotes logarithm, Δln y Δ2ln denote the first and second difference of the logarithm and lv means level.