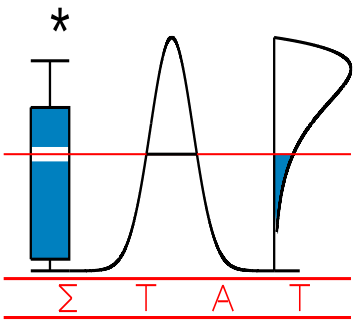
T E C H N I C A L

R E P O R T

0668

FORECASTING USING A LARGE NUMBER OF PREDICTORS : IS BAYESIAN REGRESSION A VALID ALTERNATIVE TO PRINCIPAL COMPONENTS ?

DE MOL C., GIANNONE D. and L. RIECHLIN



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Forecasting using a large number of predictors: is Bayesian regression a valid alternative to principal components?*

Christine De Mol, Université Libre de Bruxelles, ECARES
Domenico Giannone, Université Libre de Bruxelles, ECARES,
Lucrezia Reichlin, European Central Bank, ECARES and CEPR

August 2, 2006

Abstract

This paper considers Bayesian regression with normal and double-exponential priors as forecasting methods based on large panels of time series. We show that, empirically, these forecasts are highly correlated with principal component forecasts and that they perform equally well for a wide range of prior choices. Moreover, we study the asymptotic properties of the Bayesian regression under Gaussian prior under the assumption that data are quasi collinear to establish a criterion for setting parameters in a large cross-section.

JEL Classification: C11,C13,C33,C53

Keywords: Bayesian VAR, ridge regression, Lasso regression, principal components, large cross-sections

* The paper has been prepared for the conference to honor the 25th anniversary of Beveridge and Nelson's JME paper, in Atlanta March 31st-April 1st, 2006. We would like to thank Marta Banbura, James Hamilton, Christian Schumacher, Farshid Vahid, Peter Vlaar, Mark Watson for useful comments and also seminar participants at the Atlanta Federal Reserve, the 8th Bundesbank spring conference, the 5th IAP workshop, Louvain-la-Neuve, the conference on Macroeconometrics and Model Uncertainty at the Reserve Bank of New Zealand, the 2006 Australasian meeting of the Econometric Society, the 26th International Symposium on Forecasting. The opinions in this paper are those of the authors and do not necessarily reflect the views of the European Central Bank. Support by the grants "Action de Recherche Concertée" Nb 02/07-281 and IAP-network in Statistics P5/24 is gratefully acknowledged.
Please address any comments to Christine De Mol demol@ulb.ac.be; Domenico Giannone dgiannon@ulb.ac.be; or Lucrezia Reichlin lucrezia.reichlin@ecb.int.
Replication files are available at:
http://homepages.ulb.ac.be/∼dgiannon/ or http://homepages.ulb.ac.be/∼lreichli/

1 Introduction

Many problems in economics require the exploitation of large panels of time series. Recent literature has shown the "value" of large information for signal extraction and forecasting and new methods have been proposed to handle the large dimensionality problem (Forni, Hallin, Lippi, and Reichlin, 2005; Giannone, Reichlin, and Sala, 2004; Stock and Watson, 2002a,b).

A related literature has explored the performance of Bayesian model averaging for forecasting (Koop and Potter, 2003; Stock and Watson, 2004, 2005a; Wright, 2003) but, surprisingly, few papers explore the performance of Bayesian regression in forecasting with high dimensional data. Exceptions are Stock and Watson (2005) who consider normal Bayes estimators for orthonormal regressors and Giacomini and White (2006) who provide an empirical example in which a large Bayesian VAR is compared with principal component regression (PCR).

Bayesian methods are part of the traditional econometrician toolbox and offer a natural solution to overcome the curse of dimensionality problem by shrinking the parameters via the imposition of priors. In particular, the Bayesian VAR has been advocated as a device for forecasting macroeconomic data (Doan, Litterman, and Sims, 1984; Litterman, 1986). It is then surprising that, in most applications these methods have been applied to relatively small systems and that their empirical and theoretical properties for large panels have not been given more attention by the literature.

This paper is a first step towards filling this gap. We analyze Bayesian regression methods under Gaussian and double-exponential prior and study their forecasting performance on the standard "large" macroeconomic dataset that has been used to establish properties of principal component based forecast (Stock and Watson, 2002a,b). Moreover we analyze the asymptotic properties of Gaussian Bayesian regression for n, the size of the cross-section and T, the sample size, going to infinity. The aim is to establish a connection between Bayesian regression and the classical literature on forecasting with large panels based on principal components.

Our two choices for the prior correspond to two interesting cases: variable aggregation and variable selection. Under Gaussian prior, maximizing the posterior distribution generates coefficients (the mode) implying that all variables in the panel are given non-zero coefficients. Regressors, as in PCR are linear combinations of all variables in the panel, but while the Gaussian prior gives decreasing weight to the ordered eigenvalues of the covariance matrix of the data, principal components imply unit weight to the dominant ones and zero to the others. The double-exponential, on the other hand, favors *sparse* models since it puts more mass near zero and in the tails which induces a tendency of the coefficients maximizing the posterior distribution to be either large or zero. As a result, it favors the recovery of a few large coefficients instead of many small ones and truly zero rather than small values. This case is interesting because it results in variable selection rather than in variable aggregation and, in principle, this should give results that are more interpretable from the economic point of view.

Under double-exponential prior there is no analytical form for the maximizer of the posterior distribution, but we can exploit the fact that, under the prior of i.i.d. regression coefficients, the solution amounts to a Lasso (least absolute shrinkage and selection operator) regression for which there are several algorithms. In the empirics we will use two algorithms recently proposed which work without limitations of dimensionality: LARS (Least Angle Regression) developed by Efron, Hastie, Johnstone, and Tibshirani (2004) and the Iterative Landweber scheme with soft thresholding at each iteration developed by De Mol and Defrise (2002) and Daubechies, Defrise, and De Mol (2004).

An interesting feature of the Lasso regression is that it combines variable selection and parameter estimation. The estimator depends on the variable to be predicted and this may have advantages in some empirical situations. The availability of the algorithms mentioned above, which are computationally feasible, makes the double-exponential prior an attractive alternative to other priors used for variable selection such as the one proposed Fernandez, Ley, and Steel (2001) in the contest of Bayesian Model Averaging and applied by Stock and Watson (2005a) for macroeconomic forecasting with large cross-sections, which require computationally demanding algorithms.

Although Gaussian and double-exponential Bayesian regressions imply a different form of the forecast equation, an out-of-sample evaluation based on the Stock and Watson dataset, shows that, for a given range of the prior choice, the two methods produce forecasts with similar mean-square errors and which are highly correlated. These forecasts have also similar mean-square errors and are highly correlated with those produced by principal components: they do well when PCR does well. For the case of Lasso, the prior range corresponds to the selection of few variables. However, the forecasts obtained from these informative targeted predictors do not outperform PCR based on few principal components[1].

Since principal component forecasts are known to do well when variables are nearly collinear and this is a typical feature of large panels of macroeconomic data (see Giannone, Reichlin, and Sala, 2004), we study the case of Gaussian regression under near-collinearity and derive conditions on the prior parameters under which the forecast converges to the efficient one (i.e. the forecast under knowledge of the true parameters) as n, the size of the cross-section and T, the sample size, go to infinity.

The technical assumptions under which we derive the result are those that define the approximate factor structure first introduced by Chamberlain and Rothschild (1983) and generalized by Forni and Lippi (2001) and Forni, Hallin, Lippi, and Reichlin (2000). Intuitively, near-collinearity is captured by assuming that, as the size of the cross-section n increases, few eigenvalues increase while the others are bounded. Related assumptions have been introduced by Bai and Ng (2002), Bai (2003), Stock and Watson (2002a) and Stock and Watson

[1]Targeted predictors have recently found to improve performance of factor augmented forecasts when used to form principal components for factor estimation by Bai and Ng (2006). This result is not directly comparable with ours since we use targeted predictors directly as regressors.

(2002b). Bai (2003), Bai and Ng (2002), Forni, Hallin, Lippi, and Reichlin (2000), Stock and Watson (2002a) and Stock and Watson (2002b) have used them to derive the n and T asymptotic properties of the principal component regression.

This result shows how to select the prior in relation to n and helps interpreting the empirical findings. Under near-collinearity, if the prior is chosen appropriately in relation with n, Bayesian regression under normality will give weight to the principal components associated with the dominant eigenvalues and therefore will produce results which are similar to PCR. But this is what we find in the empirics which indeed shows that our data structure is nearly collinear.

However, our empirics also shows that the same results, similar performances and high correlation with PCR forecasts, are achieved by the Lasso forecast which is based on a regression on few variables. Again, we interpret this result as evidence that our panel is characterized by collinear rather than sparse covariance matrix and that few variables span the space of the pervasive common factors. These variables must be correlated with the principal components. Further work plans to explore this conjecture in more detail.

The paper is organized as follows. The second Section introduces the problem of forecasting using large cross sections. The third Section reports the result of the out-of-sample exercise for the three methods considered: principal components, Bayesian regression with normal and with double-exponential prior. The fourth Section reports asymptotic results for the (zero mean) Gaussian prior case under approximate factor structure. The fifth Section concludes and outlines problems for future research.

2 Three solutions to the "curse of dimensionality" problem

Consider the $(n \times 1)$ vector of covariance-stationary processes $Z_t = (z_{1t}, ..., z_{nt})'$. We will assume that they all have mean zero and unitary variance.

We are interested in forecasting linear transformations of some elements of Z_t using all the variables as predictors. Precisely, we are interested in estimating the linear projection

$$y_{t+h|t} = \text{proj}\left\{y_{t+h}|\Omega_t\right\}$$

where $\Omega_t = \text{span}\left\{Z_{t-p}, p = 0, 1, 2, ...\right\}$ is a potentially large time t information set and $y_{t+h} = z^h_{i,t+h} = f_h(L)z_{i,t+h}$ is a filtered version of z_{it}, for a specific i.

Traditional time series methods approximate the projection using only a finite number, p, of lags of Z_t. In particular, they consider the following regression model:

$$y_{t+h} = Z_t'\beta_0 + ... + Z_{t-p}'\beta_p + u_{t+h} = X_t'\beta + u_{t+h}$$

where $\beta = (\beta_0', ..., \beta_p')'$ and $X_t = (Z_t', ..., Z_{t-p}')'$.

Given a sample of size T, we will denote by $X = (X_{p+1}, ..., X_{T-h})'$ the $(T-h-p) \times n(p+1)$ matrix of observations for the predictors and by $y = (y_{p+1+h}, ..., y_T)'$ the $(T-h-p) \times 1$ matrix of the observations on the dependent variable. The regression coefficients are typically estimated by Ordinary Least Squares (OLS), $\hat{\beta}^{LS} = (X'X)^{-1}X'y$, and the forecast is given by $\hat{y}^{LS}_{T+h|T} = X'_T\hat{\beta}^{LS}$. When the size of the information set, n, is large, such projection involves the estimation of a large number of parameters. This implies loss of degrees of freedom and poor forecast ("curse of dimensionality problem"). Moreover, if the number of regressors is larger that the sample size, $n(p+1) > T$, the OLS is not feasible.

To solve this problem, the method that has been considered in the literature is to compute the forecast as a projection on the first few principal components (Forni, Hallin, Lippi, and Reichlin, 2005; Giannone, Reichlin, and Sala, 2004; Giannone, Reichlin, and Small, 2005; Stock and Watson, 2002a,b).

Consider the spectral decomposition of the sample covariance matrix of the regressors:

$$S_x V = VD \tag{1}$$

where $D = \text{diag}(d_1, ..., d_{n(p+1)})$ is a diagonal matrix having on the diagonal the eigenvalues of $S_x = \frac{1}{T-h-p}X'X$ in decreasing order of magnitude and $V = (v_1, ..., v_{n(p+1)})$ is the $n(p+1) \times n(p+1)$ matrix whose columns are the corresponding eigenvectors[2]. The normalized principal components (PC) are defined as:

$$\hat{f}_{it} = \frac{1}{\sqrt{d_i}} v_i' X_t \tag{2}$$

for $i = 1, \cdots, N$ where N is the number of non zero eigenvalues[3].

If most of the interactions among the variables in the information set is due to few common underlying factors, while there is limited cross-correlation among the variable specific components of the series, the information content of the large number of predictors can indeed be summarized by few aggregates, while the part not explained by the common factors can be predicted by means of traditional univariate (or low-dimensional forecasting) methods and hence just captured by projecting on the dependent variable itself (or on a small set of predictors). In such situations, few principal components, $\hat{F}_t = (\hat{f}_{1t}, ..., \hat{f}_{rt})$ with $r << n(p+1)$, provide a good approximation of the underlying factors.

Assuming for simplicity that lags of the dependent variable are not needed as additional regressors, the principal component forecast is defined as:

$$y^{PC}_{t+h|t} = \text{proj}\left\{y_{t+h}|\Omega^F_t\right\} \approx \text{proj}\left\{y_{t+h}|\Omega_t\right\} \tag{3}$$

[2]The eigenvalues and eigenvectors are typically computed on $\frac{1}{T-p}\sum_{t=p+1}^{T} X_t X_t'$ (see for example Stock and Watson, 2002a). We instead compute them on $\frac{1}{T-h-p}X'X = \frac{1}{T-p-h}\sum_{t=p+1}^{T-h} X_t X_t'$ for comparability with the other estimators considered in the paper.

[3]Note that $N \leq \min\{n(p+1), T-h-p\}$.

where $\Omega_t^F = \text{span}\left\{\hat{F}_t, \hat{F}_{t-1}, \cdots,\right\}$ is a parsimonious representation of the information set. The parsimonious approximation of the information set makes the projection feasible, since it requires the estimation of a limited number of parameters.

The literature has studied rates of convergence of the principal component forecast to the efficient forecast under assumptions defining an approximate factor structure (see the next Section). Under those assumptions, once common factors are estimated via principal components, the projection is computed by OLS treating the estimated factors as if they were observables.

The Bayesian approach consists instead in imposing limits on the length of β through priors and estimating the parameters as the posterior mode. The parameters are hence used to compute the forecasts. Here we consider two alternatives: Gaussian and double exponential prior.

Under Gaussian prior, $u_t \sim$ i.i.d. $\mathcal{N}(0, \sigma_u^2)$ and $\beta \sim \mathcal{N}(\beta_0, \Phi_0)$, and assuming for simplicity that all parameters are shrunk to zero, i.e. $\beta_0 = 0$, we have:

$$\hat{\beta}^{bay} = \left(X'X + \sigma_u^2 \Phi_0^{-1}\right)^{-1} X'y.$$

The forecast is hence computed as:

$$\hat{y}_{T+h|T}^{bay} = X_T' \hat{\beta}^{bay}.$$

In the case in which the parameters are independently and identically distributed, $\Phi_0 = \sigma_\beta^2 I$, the estimates are equivalent to those produced by penalized Ridge regression with parameter $\nu = \frac{\sigma_u^2}{\sigma_\beta^2}$[4]. Precisely[5]:

$$\hat{\beta}^{bay} = \arg\min_\beta \left\{\|y - X\beta\|^2 + \nu\|\beta\|^2\right\}.$$

There is a close relation between OLS, PCR and Bayesian regressions. For example, If the prior belief on the regression coefficients is that they are i.i.d., they can be represented as a weighted sum of the projections on the principal components:

$$X_T'\hat{\beta} = \sum_{i=1}^{N} w_i \hat{f}_{iT} \hat{\alpha}_i \tag{4}$$

where $\hat{\alpha}_i = \frac{1}{\sqrt{d_i}} v_i' X'y/(T-h-p)$ is the regression coefficient of y on the ith principal component. For OLS we have $w_i = 1$ for all i. For the Bayesian

[4]Homogenous variance and mean zero are very naive assumptions. In our case, they are justified by the fact that the variables in the panel we will consider for estimation are standardized and demeaned. This transformation is natural for allowing comparison with principal components.

[5]In what follows we will denote by $\|\cdot\|$ the L^2 matrix norm, i.e. for every matrix A, $\|A\| = \sqrt{\lambda_{max}(A'A)}$. For vectors it correspond to the Euclidean norm.

estimates $w_i = \frac{d_i}{d_i + \frac{\nu}{T-h-p}}$, where $\nu = \frac{\sigma_u^2}{\sigma_\beta^2}$. For the PCR regression we have $w_i = 1$ if $i \leq r$, and zero otherwise.

OLS, PCR and Gaussian Bayesian regression weight all variables. An alternative is to select variables. For Bayesian regression, variable selection can be achieved by a double exponential prior, which, when coupled with a zero mean i.i.d. prior, is equivalent to the method that is sometimes called Lasso regression (least absolute shrinkage and selection operator)[6]. In this particular i.i.d. prior case the method can also be seen as a penalized regression with a penalty on the coefficients involving the L_1 norm instead of the L_2 norm. Precisely:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \nu \sum_{i=1}^{n} |\beta_i| \right\} \tag{5}$$

where $\nu = \frac{1}{\tau}$ where τ is the scale parameter of the prior density[7] (see e.g. Tibshirani, 1996).

Compared with the Gaussian density, the double-exponential puts more mass near zero and in the tails and this induces a tendency to produce estimates of the regression coefficients that are either large or zero. As a result, one favors the recovery of a few large coefficients instead of many fairly small ones. Moreover, as we shall see, the double-exponential prior favors truly zero values instead of small ones, i.e. it favors *sparse* regression coefficients (sparse mode).

To gain intuition about Lasso regression, let us consider, as an example, the case of orthogonal regressors, a case for which the posterior mode has known analytical form. In particular, let us consider the case in which the regressors are the principal components of X. In this case, Lasso has the same form of (4) with $w_i\hat{\alpha}_i$ replaced by $S_\nu(\hat{\alpha}_i)$ where S_ν is the *soft-thresholder* defined by

$$S_\nu(\alpha) = \begin{cases} \alpha + \nu/2 & \text{if} \quad \alpha \leq -\nu/2 \\ 0 & \text{if} \quad |\alpha| < \nu/2 \\ \alpha - \nu/2 & \text{if} \quad \alpha \geq \nu/2. \end{cases} \tag{6}$$

As seen, this sparse solution is obtained by setting to zero all coefficients $\hat{\alpha}_i$ which in absolute value lie below the threshold $\nu/2$ and by shrinking the largest ones by an amount equal to the threshold. Let us remark that it would also be possible to leave the largest components untouched, as done in so-called *hard-thresholding*, but we do not consider this variant here since the lack of continuity of the function $S_\nu(\alpha)$ makes the theoretical framework more complicated.

In the general case, i.e. with non orthogonal regressors, the Lasso solution will enforce sparsity on the variables themselves rather than on the principal components and this is an interesting feature of the method since it implies a regression on few observables rather than on few linear combinations of the observables. Note that the model with non-Gaussian priors is not invariant under orthogonal linear transformation of the data.

[6]It should be noted however that Lasso is actually the name of an algorithm for finding the maximizer of the posterior proposed in Tibshirani (1996).

[7]We recall here that the variance of the prior density is proportional to $2\tau^2$.

Notice also that, unlike Ridge and PCR, where the selection of the regressors is performed independently of the choice of the series to be forecasted, the Lasso regression depends on that choice.

Methods described by equation (4) will perform well provided that no truly significant coefficients α_i are observed for $i > r$, because in principal component regression they will not be taken into account and in Ridge their influence will be highly weakened. Bad performances are to be expected if, for example, we aim at forecasting a time series y_t, which by bad luck is just equal or close to a principal component $\hat{f}_i$ with $i > r$. Lasso solves this problem.

Unfortunately, in the general case the maximizer of the posterior distribution has no analytical form and has to be computed using numerical methods such as the Lasso algortithm of Tibshirani (1996) or quadratic programming based on interior point methods advocated in Chen, Donoho, and Saunders (2001). Two efficient alternatives to the Lasso algorithm, which work without limitations of dimensionality also for sample size T smaller than the number of regressors $n(p+1)$, have been developed more recently by Efron, Hastie, Johnstone, and Tibshirani (2004) under the name LARS (Least Angle Regression)[8] and by De Mol and Defrise (2002); Daubechies, Defrise, and De Mol (2004) who use instead an Iterative Landweber scheme with soft thresholding at each iteration[9].

The next section will consider the empirical performance of the three methods discussed in an out-of-sample forecast exercise based on a large panel of time series.

3 Empirics

The data set employed for the out-of-sample forecasting analysis is the same as the one used in Stock and Watson (2005b). The panel includes real variables (sectoral industrial production, employment and hours worked), nominal variables (consumer and producer price indices, wages, money aggregates), asset prices (stock prices and exchange rates), the yield curve and surveys. A full description is given in Appendix C.

Series are transformed to obtain stationarity. In general, for real variables, such as employment, industrial production, sales, we take the monthly growth rate. We take first differences for series already expressed in rates: unemployment rate, capacity utilization, interest rate and some surveys. Prices and wages are transformed to first differences of annual inflation following Giannone, Reichlin, and Sala (2004); Giannone, Reichlin, and Small (2005).

Let us define IP as the monthly industrial production index and CPI as the consumer price index. The variables we forecast are

[8]The LARS algorithm has also been used in econometric forecasting by Bai and Ng (2006) for selecting variables to form principal components in factor augmented forecasts.

[9]The latter algorithm carries out most of the intuition of the orthogonal regression case and is described in Appendix B. For the LARS algorithm we refer to Efron, Hastie, Johnstone, and Tibshirani (2004).

$$z^h_{IP,t+h} = (ip_{t+h} - ip_t) = z_{IP,t+h} + ... + z_{IP,t+1}$$

and

$$z^h_{CPI,t+h} = (\pi_{t+h} - \pi_t) = z_{CPI,t+h} + ... + z_{CPI,t+1}$$

where $ip_t = 100 \times \log IP_t$ is the (rescaled) log of IP and $\pi_t = 100 \times \log \frac{CPI_t}{CPI_{t-12}}$ is the annual CPI inflation (IP enters in the pre-transformed panel in first log differences, while annual inflation in first differences).

The forecasts for the (log) IP and the level of inflation are recovered as:

$$\widehat{ip}_{T+h|T} = \hat{z}^h_{IP,T+h|T} + ip_T$$

and

$$\widehat{\pi}_{T+h|T} = \hat{z}^h_{CPI,T+h|T} + \pi_T$$

The accuracy of predictions is evaluated using the mean-square forecast error ($MSFE$) metric, given by:

$$MSFE^h_\pi = \frac{1}{T_1 - T_0 - h + 1} \sum_{T=T_0}^{T_1-h} (\widehat{\pi}_{T+h|T} - \pi_{T+h})^2$$

and

$$MSFE^h_{ip} = \frac{1}{T_1 - T_0 - h + 1} \sum_{T=T_0}^{T_1-h} (\widehat{ip}_{T+h|T} - ip_{T+h})^2$$

The sample has a monthly frequency and ranges from 1959:01 to 2003:12. The evaluation period is 1970:01 to 2002:12. T_1=2003:12 is the last available point in time, T_0= 1969:12 and $h = 12$. We consider rolling estimates with a window of 10 years, i.e. parameters are estimated at each time T using the most recent 10 years of data.

All the procedures are applied to standardized data. Mean and variance are re-attributed to the forecasts accordingly.

We report results for industrial production (IP) and the consumer price index (CPI).

Let us start from principal component regression. We report results for the choice of $r = 1, 3, 5, 10, 25, 50, 75$ principal components. The case $r = 0$ is the forecast implied from a random walk with drift on the log of IP and the annual CPI inflation, while $r = n$ is the OLS solution. We only report results for $p = 0$ which is the one typically considered in macroeconomic applications and for which the theory has been developed[10] .

We report MSFE relative to the random walk, and the variance of the forecasts relative to the variance of the series of interest. The MSFE is also reported

[10]The empirical literature has also consider the inclusion of the past of the variable of interest to capture series specific dynamic. We do not consider this case here since for once few PC are included, series specific dynamics does not help forecasting our variables of interest as shown in D'Agostino and Giannone (2005)

for two sub-samples: the first half of the evaluation period 1970-1985, and the second half 1985-2002. This would help us understand the relative performance of the methods also in a case where the predictability of key macroeconomic time series has dramatically decreased (on this point, see D'Agostino, Giannone, and Surico (2006)). Results are reported in Table 1.

Table 1: Principal component forecasts

Industrial Production							
	Number of Principal Components						
	1	3	6	10	25	50	75
MFSE 1971-2002	0.91	0.62	0.56	0.54	0.65	0.93	1.56
MFSE 1971-1984	0.89	0.45	0.35	0.34	0.46	0.70	1.18
MFSE 1985-2002	0.98	1.13	1.16	1.13	1.21	1.60	2.68
Variance*	0.23	0.70	0.79	0.97	1.28	1.43	1.78

Consumer Price Index							
	Number of Principal Components						
	1	3	6	10	25	50	75
MFSE 1971-2002	0.57	0.55	0.57	0.69	0.83	1.17	1.69
MFSE 1971-1984	0.48	0.40	0.39	0.48	0.56	0.89	1.23
MFSE 1985-2002	1.03	1.28	1.43	1.71	2.11	2.47	3.83
Variance*	0.36	0.55	0.61	0.63	0.69	0.89	1.69

MSFE are relative to a the Naive, Random Walk, forecast. *The variance of the forecast relative to the variance of the series.

Let us start from the whole evaluation sample. Results show that principal components improve a lot over the random walk both for IP and CPI. The advantage is lost when taking too many PC, which implies loss of parsimony. Notice that, as the number of PC increases, the variance of the forecasts becomes larger to the point of becoming larger than the variance of the series itself. This is explained by the large sample uncertainty of the regression coefficients when there is a large number of regressors. Looking at the two sub-samples, we see that PCs perform very well in the first part of the sample, while in the most recent period they perform very poorly, worse than the random walk.

For comparability, we focus on the case $p = 0$ also for the Bayesian regression (no lags of the regressor). Note, that, for $h = 1$, this case corresponds to a row of a VAR of order one. The exercise is for the i.i.d. Gaussian prior (Ridge regression). This prior works well for the $p = 0$ case considered here. However, for the case $p > 0$, it might be useful to shrink more the coefficients of additional lagged regressors, as, for example, with the Minnesota prior (Doan, Litterman, and Sims, 1984; Litterman, 1986). This is beyond the scope of the present empirical analysis which is meant as a first assessment of the general

performance of the methods[11].

For the Bayesian (Ridge) case, we run the regression using the first estimation sample 1959-1969 for a grid of priors. We then choose the priors for which the in-sample fit explains a given fraction $1-\kappa$ of the variance of the variable to be forecast. We report results for different values of κ (the associated ν, which are kept fixed for the whole out-of-sample evaluation period, are also reported). Notice that $\kappa = 1$ corresponds to the random walk since, in this case, all coefficients are set to zero. The other extreme, κ close to 0, is associated with a quite uninformative prior and hence will be very close to the OLS. Results are reported in Table 2.

Table 2: Bayesian forecasts with Gaussian prior

Industrial Production									
	In-sample Residual variance								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν	6	25	64	141	292	582	1141	2339	6025
MFSE 1971-2002	0.96	0.70	0.60	0.56	0.56	0.58	0.64	0.72	0.83
MFSE 1971-1984	0.74	0.50	0.41	0.38	0.40	0.44	0.52	0.63	0.78
MFSE 1985-2002	1.59	1.31	1.16	1.08	1.03	1.00	0.98	0.98	0.98
Variance*	0.71	0.63	0.57	0.49	0.39	0.29	0.19	0.12	0.07
Correlation with PC forecasts (r=10)	0.62	0.81	0.89	0.92	0.93	0.91	0.85	0.74	0.48

Consumer Price Index									
	In-sample Residual variance								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν	16	60	143	288	528	949	1751	3532	9210
MFSE 1971-2002	0.88	0.72	0.66	0.63	0.62	0.63	0.66	0.73	0.84
MFSE 1971-1984	0.72	0.58	0.52	0.51	0.51	0.54	0.59	0.68	0.82
MFSE 1985-2002	1.60	1.41	1.29	1.19	1.11	1.04	0.98	0.95	0.95
Variance*	0.41	0.35	0.32	0.28	0.24	0.19	0.13	0.08	0.05
Correlation with PC forecasts (r=10)	0.68	0.86	0.92	0.94	0.92	0.89	0.83	0.69	0.33

MSFE are relative to a the Naive, Random Walk, forecast. *The variance of the forecast relative to the variance of the series.

The Ridge forecast performs well for a range of κ between 30% and 70% that are associated with shrinkage parameters between half and ten times the cross-sectional dimension, n. For the whole sample, the MSFE are close to that obtained with principal component regression. Moreover, the forecasts produced by Ridge regressions are smoother than the PC forecasts, which is a desirable property.

[11] An additional feature of the Litterman priors is to shrink less coefficients associated with the variable we are interested in forecasting. This can be helpful when series specific dynamics have significant forecasting power. We do not consider here this case for comparability with the PC case. See Footnote 3.

The last line of the table shows the correlation among Ridge forecasts and principal component forecasts[12]. Principal components and Ridge forecasts are highly correlated, particularly when the prior is such that the forecasting performances are good. The fact that correlation is maximal for parameters giving the best forecasts suggests that there is a common explanation for the good performance of the two methods.

As for the two sub-samples, results are also qualitatively similar to principal component forecasts. Ridge performs particularly well in the first sub-sample but loses all the advantage in the second. We can note, however, more stability than in the principal components case. This is not surprising since Ridge uses all eigenvalues in decreasing importance instead of truncating after r as in the principal components case. Notice also that, for inflation, with ν in the intermediate range, even in the most recent sample there is a slight improvement over the random walk.

Finally, we analyze the case of double-exponential priors. In this case, instead of fixing the values of the parameter ν, we select the prior that delivers a given number (k) of non zero coefficients at each estimation step in the out-of-sample evaluation period. We look at the cases of $k = 1, 3, 5, 10, 25, 50, 75$ non-zero coefficients[13].

Results, reported in Table 3, show that good forecasts are obtained with a limited number of predictors, between 5 and 25. As for Ridge, maximal correlation with the principal component forecast is achieved for the selection of parameters that gives the best results.

Comparable MSFE for the three methods as well as the correlation of the forecast suggest that the covariance of our data are characterized by few dominant eigenvalues. In this case, both PC and Ridge, by keeping the largest ones and giving, respectively zero weight and small weight to the others, should perform similarly. This point will emerge more clearly in next Section on the basis of the asymptotic analysis.

The result for Lasso is less straightforward to interpret since this is a regression on few variables rather than on few aggregates of the variables. The high correlation of the Lasso forecast with the PC forecast implies two things. First, the panel must be characterized by collinearity rather than sparsity and, second, few variables must span approximately the space of the pervasive common factors.

Again, the correlation of the two Bayesian forecasts with the principal component forecast, for the priors that ensure good performance, implies that there must be a common explanation for the success of the three methods.

[12]For the principal component forecasts we use $r = 10$. We obtain similar results also for $r = 3, 5$, i.e. when PC forecasts perform well.

[13]For this exercise we use the LARS algorithm which delivers at once the Lasso solutions for any given number of non zero coefficients, for $k = 1, ..., n$. An alternative is to select the prior ν that deliver a given number (k) of non zero coefficients in the initial sample $1959 - 1970$. Then the prior ν can be kept fixed at each estimation step as done for the Ridge case. In this case, we can use the iterated Landweber algorithm with soft thresholding whose input is the prior ν rather than the number of non-zero coefficients. This alternative strategy provides qualitatively similar results. They are available on request.

Table 3: Lasso forecasts

Industrial Production							
	Number of non-zero coefficients						
	1	3	5	10	25	50	60
MFSE 1971-2002	0.86	0.69	0.64	0.60	0.64	0.77	1.10
MFSE 1971-1984	0.80	0.56	0.50	0.44	0.47	0.58	0.91
MFSE 1985-2002	1.05	1.05	1.05	1.07	1.14	1.32	1.67
Variance*	0.07	0.16	0.24	0.40	0.53	0.65	0.79
Correlation with PC forecasts (r=10)	0.05	0.64	0.81	0.85	0.84	0.68	0.44

Consumer Price Index							
	Number of non-zero coefficients						
	1	3	5	10	25	50	75
MFSE 1971-2002	0.90	0.76	0.62	0.59	0.68	0.86	1.06
MFSE 1971-1984	0.88	0.70	0.54	0.48	0.52	0.70	0.93
MFSE 1985-2002	1.00	1.04	1.02	1.14	1.44	1.65	1.68
Variance*	0.05	0.09	0.18	0.26	0.33	0.39	0.50
Correlation with PC forecasts (r=10)	0.05	0.64	0.81	0.85	0.84	0.68	0.44

MSFE are relative to a the Naive, Random Walk, forecast. *The variance of the forecast relative to the variance of the series.

The variables selected for $k \approx 10$ at the beginning and at the end of the out-of-sample evaluation period are reported in the last two column of the table describing the database in Appendix C. Two main results emerge. First, only some of the variables selected are those typically included in small-medium size models: the commodity price indexes, the spreads, money aggregates and stock market variables. Some of the selected variables are sectoral (production, labor market and price indicators) or regional (housing). Second, the selection is different at different points in the sample. Only one variable selected at the beginning of the 70s is also picked-up in the most recent period for CPI inflation forecasts. For IP forecasts, no variables are selected in both periods.

We have two conjectures about these results. The fact that variables are not clearly interpretable probably indicates that the panel contains clusters of correlated variables and the procedure selects a particular one, not necessarily the most meaningful from the economic point of view. This implies that variable selection methods are not easily interpretable in this case. The fact that the procedure selects different variables at different points of the sample, implies temporal instability. On the other hands, results also indicate that temporal instability does not affect the relative performance of principal components and Ridge with respect to Lasso. This suggests that principal components and Ridge, by aggregating all variables in the panel, stabilize results providing a sort of insurance against temporal instability. These conjectures will be explored in

further work.

4 Theory

We have seen that the Bayesian regression and PCR regression can be seen as ways of stabilizing OLS when data are nearly collinear (sometimes called regularization methods). Large panels of macroeconomic time series are typically highly collinear (Giannone, Reichlin, and Sala, 2004) so that these methods are also appropriate to deal with the "curse of dimensionality" problem.

This observation motivates the assumptions that we will now introduce to define the asymptotic analysis.

A1) X_t has the following representation[14]:

$$X_t = \Lambda F_t + \xi_t$$

where $F_t = (f_{1t}, ..., f_{rt})'$, the common factors, is an r-dimensional stationary process with covariance matrix $\mathrm{E}F_tF_t' = I_r$ and ξ_t, the idiosyncratic components, is an $n(p+1)$-dimensional stationary process with covariance matrix $\mathrm{E}\xi_t\xi_t' = \Psi$.

A2) $y_{t+h} = \gamma F_t + v_{t+h}$ where v_{t+h} is orthogonal to F_t and ξ_t.

Assumption A1 can be understood as a quasi-collinearity assumption whereby the bulk of cross-correlation is driven by few orthogonal common factors while the idiosyncratic components are allowed to have a limited amount of cross-correlation. The conditions that limit the cross-sectional correlation are given below (condition CR2).

Under assumption A2, if the common factors F_t were observed, we would have the unfeasible optimal forecast[15]:

$$y^*_{t+h|t} = \gamma F_t$$

Following Forni, Hallin, Lippi, and Reichlin (2000, 2005); Forni, Giannone, Lippi, and Reichlin (2005), we will impose two sets of conditions, conditions that ensure stationarity (see appendix A) and conditions on the cross-sectional correlation as n increases[16]. These conditions are a generalization to the dynamic case of the conditions defining an approximate factor structure given by Chamberlain and Rothschild (1983). Precisely:

[14]Notice that here we define the factor model over $X_t = (Z_t', ...Z_{t-p}')'$ while the literature typically defines it over Z_t. It can be seen that if Z_t follows an approximate factor structure defined below, with k common factors, then also X_t follows an approximate factor structure with $r \leq k(p+1)$ common factors.

[15]We are assuming here that common factors are the only source of forecastable dynamics. We make this assumption for simplicity and since from an empirical point of view series specific dynamics does not help forecasting our variables of interest. See footnote 3.

[16]Bai (2003), Bai and Ng (2002) and Stock and Watson (2002a) give similar conditions.

CR1) $0 < \liminf_{n\to\infty} \frac{1}{n}\lambda_{min}\left(\Lambda'\Lambda\right) < \limsup_{n\to\infty} \frac{1}{n}\lambda_{max}\left(\Lambda'\Lambda\right) < \infty$

CR2) $\limsup_{n\to\infty} \lambda_{max}\left(\Psi\right) < \infty$ and $\liminf_{n\to\infty} \lambda_{min}\left(\Psi\right) > 0$

Note that $CR1$ implies that as the cross-sectional dimensional increases few eigenvalues of $\Sigma_x = \Lambda\Lambda' + \Psi$ remain pervasive while $CR2$ implies that the others are asymptotically bounded.

We study now the properties of the Bayesian estimates if the data are generated from an approximate factor structure. Let us first notice that under our assumptions we have $\Sigma_x = \mathrm{E}(X_t X_t') = (\Lambda\Lambda' + \Psi)$ and $\Sigma_{xy} = \mathrm{E}(X_t y_{t+h}) = \Lambda\gamma'$. Consequently, the population regression coefficients are given by

$$\beta = \Sigma_x^{-1}\Sigma_{xy} = (\Lambda\Lambda' + \Psi)^{-1}\Lambda\gamma' = \Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\gamma'$$

Consider now

$$y_{t+h|t} = X_t'\beta = F_t\Lambda'\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\gamma' + \xi_t\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\gamma'$$

.

Under assumptions CR1-2 we have

$$\Lambda'\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1} = I + O\left(\frac{1}{n}\right)$$

since

$$\begin{aligned}\|\Lambda'\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1} - I\| &= \|\Lambda'\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}(\Lambda'\Psi^{-1}\Lambda)^{-1}\| \\ &\leq \|(\Lambda'\Psi^{-1}\Lambda)^{-1}\|^2\|\Lambda'\Psi^{-1}\Lambda\| \leq \left(\tfrac{\lambda_{max}(\Psi)}{\lambda_{min}(\Lambda'\Lambda)}\right)^2 \tfrac{\lambda_{max}(\Lambda'\Lambda)}{\lambda_{min}(\Psi)} = O\left(\tfrac{1}{n}\right)\end{aligned}$$

Moreover, $\beta = O\left(\frac{1}{\sqrt{n}}\right)$ since

$$\|\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\| \leq \frac{\lambda_{max}(\Psi)}{\lambda_{min}(\Lambda'\Lambda)}\sqrt{\frac{\lambda_{max}(\Lambda'\Lambda)}{\lambda_{min}(\Psi)}} = O\left(\frac{1}{\sqrt{n}}\right)$$

This implies that:

$$\mathrm{E}\left[(\xi_t\beta)^2\right] = \beta'\Psi\beta \leq \|\beta\|^2\|\Psi\| = O\left(\frac{1}{n}\right)$$

hence by the Markov's inequality we have:

$$\xi_t\beta = O_p\left(\frac{1}{\sqrt{n}}\right)$$

This proves the following result:

Proposition 1 Under assumptions A1-2 and CR1-2 we have:

$$y_{t+h|t} = \gamma F_t + O_p\left(\frac{1}{\sqrt{n}}\right) \text{ as } n \to \infty$$

The result above tells us that, under the factor model representation, the projection over the whole dataset X_t and the projection over the unobserved common factors F_t are asymptotically equivalent for $n \to \infty$.

In Proposition 1 we assume that the second order moments of the data are known when performing the projection. What if they are estimated? Under the above Assumptions, it has been shown that the forecasts based on the regression on the first r principal components provide consistent estimates for $y^*_{t+h|t}$. The Proposition below gives conditions for the shrinkage parameter that allow to obtain consistent forecasts from Bayesian regression under Gaussian priors. We will need the additional Assumption A3 that insures that the elements of the sample covariances of X_t and y_t converge uniformly to their population counterpart, see the Appendix A for details.

Proposition 2 Under assumptions A1-2 and CR1-2, if $\liminf_{n\to\infty} \frac{\lambda_{min}(\Phi_0)}{\|\Phi_0\|} > 0$ then:

$$X_t'\hat{\beta}^{bay} = X_t'\beta + O_p\left(\frac{1}{nT\|\Phi_0\|}\right) + O_p\left(n\sqrt{T}\|\Phi_0\|\right) \text{ as } n, T \to \infty,$$

provided that $\frac{1}{nT}\|\Phi_0\|^{-1} \to 0$ and $\frac{1}{n\sqrt{T}}\|\Phi_0\|^{-1} \to \infty$ as $n, T \to \infty$,

Proof. See the Appendix.

If coefficients are i.i.d. $\mathcal{N}(0, \sigma_\beta^2)$, then the conditions are satisfied if $\sigma_\beta^2 = \frac{1}{cnT^{1/2+\delta}}$, where c is an arbitrary positive constant. Hence, we should shrink the single regressors with an asymptotic rate faster than the $\frac{1}{n}$. With non i.i.d. prior, the condition $\liminf_{n\to\infty} \frac{\lambda_{min}(\Phi_0)}{\|\Phi_0\|} > 0$ requires that all the regression coefficients should be shrunk at the same asymptotic rate.

Combining Propositions 1 and 2 we obtain:

Corollary Under the assumptions A1-2 and CR1-2 and provided the conditions of Proposition 2 are satisfied, we have

$$X_t'\hat{\beta}^{bay} = \gamma F_t + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{nT\|\Phi_0\|}\right) + O_p\left(n\sqrt{T}\|\Phi_0\|\right) \text{ as } n, T \to \infty.$$

A suitable choice for the prior is $\|\Phi_0\| = \frac{1}{cnT^{1/2+\delta}}$. In this case we have:

$$\Delta_{nT}\left(X_t'\hat{\beta}^{bay} - \gamma F_t\right) = O_p(1) \text{ as } n, T \to \infty,$$

where $\Delta_{nT} = \min\left\{\sqrt{n}, T^{\delta}, T^{(\frac{1}{2}-\delta)}\right\}$ and $0 < \delta < 1/2$. These rates of consistency are different from the ones derived for principal components in Forni, Giannone, Lippi, and Reichlin (2005) and, using a different set of assumptions by Bai (2003), and probably can be improved by imposing further assumptions.

Proposition 2 tells us that, under the factor structure assumption, the Bayesian regression should use a prior that, as the cross-section dimension increases, shrinks increasingly more all regression coefficients to zero. The reason is that, if the factors are pervasive in the sense of condition CR1, then all variables are informative for the common factors and we should give weight to all of them. Consequently, as the number of predictors increases, the magnitude of each regression coefficient has to decrease.

The intuition of this result is very simple. The factor structure implies that there are few r dominant eigenvalues that diverge faster than the remaining smaller ones as the cross-section dimension increases. The parameter's prior chosen as above ensures that the effect of the factors associated with the dominant eigenvalues is not distorted asymptotically while for the effect of the smaller ones goes to zero asymptotically. Clearly, as mentioned in the empirical Section, if there are few dominant eigenvalues, both Bayesian regression under normality and PCR will only give weight to the principal components associated to the dominant eigenvalues.

5 Conclusions and open questions

This paper has analyzed the properties of Bayesian regression in large panels of time series and compared them to PCR.

We have considered the Gaussian and the double exponential prior and show that they offer a valid alternative to principal components. For the macroeconomic panel considered, the forecast they provide is very correlated to that of PCR and implies similar mean-square forecast errors.

This exercise should be understood as rather stylized. For the Bayesian case there is room for improvement, in particular by using developments in BVAR (Doan, Litterman, and Sims, 1984; Litterman, 1986) and related literature.

In the asymptotic analysis we have considered the Gaussian prior case. For that case, we have shown n, T rates of convergence to the efficient forecast under an approximate factor structure. This analysis guides us in the setting of the prior, also interpreted as a Ridge penalization parameter. The empirical analysis reports results for the optimal parameter and for a larger range of parameter choice. The setting of the parameters for the double-exponential case has been exclusively empirical. It is designed so as to deliver a given number of non zero coefficients at each estimation step in the out-of-sample evaluation period. The algorithm provides good results by selecting few variables in the regression.

These results show that our data, which correspond to the typical macroeconomic data-set used for macroeconomic policy analysis, is characterized by collinearity rather than sparsity. On the other hand, the result that few selected variables are able to capture the space spanned by the common factors,

suggests that small models with accurately selected variables may do as well as methods that use information on large panels and are based on regressions on linear combinations of all variables. This point calls for further research since our results show that the variable selection provided by the Lasso regression is not clearly interpretable and they are not the typical ones that a macroeconomist would include in a VAR. Moreover, the selected variables change over time. A conjecture, to be explored in further work, is that, although the underlying model implies parameter instability, a well chosen approximating model based on a large cross-section has a chance of performing well in forecast since the use of a large number of variables works as a sort of insurance against parameter instability.

References

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71(1), 135–171.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

——— (2006): "Forecasting Economic Time Series Using Targeted Predictors," Manuscript, University if Michigan.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, factor structure and mean-variance analysis in large asset markets.," *Econometrica*, 51, 1305–1324.

Chen, S. S., D. Donoho, and M. Saunders (2001): "Atomic Decomposition by Basis Pursuit," *SIAM Review*, 43, 129–159.

D'Agostino, A., and D. Giannone (2005): "Comparing Alternative Predictors Based on Large-Panel Factor Models," Manuscript, Université Libre de Bruxelles.

D'Agostino, A., D. Giannone, and P. Surico (2006): "(Un)Predictability and Macroeconomic Stability," Working Paper Series 605, European Central Bank.

Daubechies, I., M. Defrise, and C. De Mol (2004): "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, 57, 1416–1457.

De Mol, C., and M. Defrise (2002): "A note on wavelet-based inversion methods," in *Inverse Problems, Image Analysis and Medical Imaging*, ed. by M. Z. Nashed, and O. Scherzer, pp. 85–96. American Mathematical Society.

Doan, T., R. Litterman, and C. A. Sims (1984): "Forecasting and Conditional Projection Using Realistic Prior Distributions," *Econometric Reviews*, 3, 1–100.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): "Least angle regression," *Ann. Statist.*, 32, 407–499.

Fernandez, C., E. Ley, and M. F. J. Steel (2001): "Benchmark priors for Bayesian model averaging," *Journal of Econometrics*, 100(2), 381–427.

Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2005): "Opening the Black Box: Structural Factor Models with large cross-sections," Manuscript, Université Libre de Bruxelles.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Dynamic Factor Model: identification and estimation," *Review of Economics and Statistics*, 82, 540–554.

——— (2005): "The Generalized Dynamic Factor Model: one-sided estimation and forecasting," *Journal of the American Statistical Association*, 100, 830–840.

Forni, M., and M. Lippi (2001): "The Generalized Dynamic Factor Model: representation theory.," *Econometric Theory*, 17, 1113–1141.

Giacomini, R., and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, forthcoming.

Giannone, D., L. Reichlin, and L. Sala (2004): "Monetary Policy in Real Time," in *NBER Macroeconomics Annual*, ed. by M. Gertler, and K. Rogoff, pp. 161–200. MIT Press.

Giannone, D., L. Reichlin, and D. Small (2005): "Nowcasting GDP and inflation: the real-time informational content of macroeconomic data releases," Finance and Economics Discussion Series 2005-42, Board of Governors of the Federal Reserve System (U.S.).

Koop, G., and S. Potter (2003): "Forecasting in large macroeconomic panels using Bayesian Model Averaging," Staff Reports 163, Federal Reserve Bank of New York.

Litterman, R. (1986): "Forecasting With Bayesian Vector Autoregressions – Five Years of Experience," *Journal of Business and Economic Statistics*, 4, 25–38.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 147–162.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes.," *Journal of Business and Economics Statistics*, 20, 147–162.

Stock, J. H., and M. W. Watson (2004): "Forecasting with many predictors," Unpublished manuscript, Princeton University.

Stock, J. H., and M. W. Watson (2005a): "An Empirical Comparison Of Methods For Forecasting Using Many Predictors," Manuscript, Princeton University.

Stock, J. H., and M. W. Watson (2005b): "Implications of Dynamic Factor Models for VAR Analysis," Unpublished manuscript, Princeton University.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, 58, 267–288.

Wright, J. H. (2003): "Forecasting U.S. inflation by Bayesian Model Averaging," International Finance Discussion Papers 780, Board of Governors of the Federal Reserve System (U.S.).

6 Appendix A: Proof of Proposition 2

Denote:

- by y_t the generic variable to be forecast as $y_t = z_{it}^h$
- the covariance matrix of the regressors as $\Sigma_x = \mathrm{E}(x_t x_t')$. The sample equivalent will be denoted by $S_x = X'X/T$. The estimation error will be denote by $E_x = \Sigma_x - S_x$. These matrices are of dimension $n \times n$.
- the covariance matrix of the regressors and the variable to be predicted as $\Sigma_{xy} = \mathrm{E}(x_t y_{t+h}')$. The sample equivalent will be denoted by $S_{xy} = X'y/T$. The estimation error will be denote by $E_{xy} = \Sigma_{xy} - S_{xy}$. These matrices are of dimension $n \times 1$.

We assume stationarity.
Moreover, we need the following assumption:

A3) There exists a positive constant $K \leq \infty$, such that for all $T \in \mathbb{N}$ and $i, j \in \mathbb{N}$

$$T\ \mathrm{E}[(e_{x,ij})^2] < K \quad \text{and } T\ \mathrm{E}[(e_{xy,i})^2] < K$$

as $T \to \infty$, where $e_{x,ij}$ denote the i,jth entry of E_x and $e_{xy,i}$ denote the ith entry of E_{xy}. Sufficient conditions can be found in Forni, Giannone, Lippi, and Reichlin (2005).

Remark 1 We can consider here without loss of generality the case of iid prior on the coefficients and we will denote by $\tilde{\nu} = \frac{\sigma_u^2}{T\|\Phi_0\|}$ the rescaled penalization in the Ridge regression. In fact, in the case of non-iid prior, we can redefine

the regression in terms of $\tilde{Z}_t = \frac{\Phi_0^{1/2} Z_t}{\sqrt{\|\Phi_0\|}}$. Then the resulting regression coefficients, $\tilde{\beta} = \sqrt{\|\Phi_0\|}\Phi_0^{-1/2}\beta$ will be iid with prior variance $\|\Phi_0\|$. Moreover the transformed regressors $\tilde{Z}_t$ have the factor representation

$$\tilde{Z}_t = \tilde{\Lambda} F_t + \tilde{\xi}_t$$

where $\tilde{\Lambda} = \frac{\Phi_0^{1/2}\Lambda}{\sqrt{\|\Phi_0\|}}$ and $\tilde{\xi}_t = \frac{\Phi_0^{1/2}\xi_t}{\sqrt{\|\Phi_0\|}}$. The assumption $\liminf_{n\to\infty} \frac{\lambda_{min}(\Phi_0)}{\|\Phi_0\|} > 0$ insures that the transformed model still satisfies conditions CR1 and CR2.

Remark 2 In what follows we will prove the proposition for the case $p = 0$. The case $p > 0$ can be analyzed similarly by noticing that if Z_t possesses an approximate factor structure then also X_t has it.

Defining $\Sigma_x(\tilde{\nu}) = \Sigma_x + \tilde{\nu} I_n$ and the sample equivalent $S_x(\tilde{\nu}) = S_x + \tilde{\nu} I_n$, we are interested in the properties of $\beta(\tilde{\nu})$ and $\hat{\beta}(\tilde{\nu})$ which are solutions of the following linear system of equations:

$$\begin{array}{c} \Sigma_x(\tilde{\nu})\beta(\tilde{\nu}) = \Sigma_{xy} \\ S_x(\tilde{\nu})\hat{\beta}(\tilde{\nu}) = S_{xy} \end{array} \tag{7}$$

Notice that $\beta(0) = \beta$ is the population regression coefficient and $\hat{\beta}(0) = \hat{\beta}$ is the sample regression coefficient. For $\tilde{\nu} > 0$ we have the Ridge regression coefficients.

Lemma 1 Under assumptions CR1-2 we have

$$\|\beta(\tilde{\nu})\| = O\left(\frac{1}{\sqrt{n}}\right) \tag{8}$$

and

$$\|\beta - \beta(\tilde{\nu})\| = O\left(\tilde{\nu}\ n^{-3/2}\right) \text{ as } n \to \infty\ . \tag{9}$$

Proof. We have:

$$\|\beta(\tilde{\nu})\| = \|(\Lambda\Lambda' + \Psi + \tilde{\nu} I_n)^{-1}\Lambda\gamma'\| \leq \|(\Lambda\Lambda')^{-1}\|\|\Lambda\|\|\gamma'\|$$

which implies (8) by assumption CR1.

On the other hand, recalling that $\beta = \beta(0)$, we have

$$\begin{array}{rl} \beta - \beta(\tilde{\nu}) = & \left[(\Lambda\Lambda' + \Psi)^{-1} - (\Lambda\Lambda' + \Psi + \tilde{\nu} I_n)^{-1}\right]\Lambda\gamma' \\ & = (\Lambda\Lambda' + \Psi)^{-1}\tilde{\nu} I_n (\Lambda\Lambda' + \Psi + \tilde{\nu} I_n)^{-1}\Lambda\gamma' \end{array}$$

thanks to the matrix identity

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}\ . \tag{10}$$

Hence

$$\|\beta - \beta(\tilde{\nu})\| \leq \tilde{\nu}\|\Lambda\Lambda'\|^{-2}\|\Lambda\| \ \|\gamma'\| = O(\tilde{\nu}\ n^{-3/2})$$

by assumption CR1. Q.E.D.

Whereas for $\tilde{\nu} = 0$ the optimal regression coefficient β provides consistent forecasts, the Ridge parameter $\tilde{\nu}$ introduces a bias which tends to zero for large cross-sectional dimensions provided that it does not increase too fast relatively to the cross-sectional dimension n. Let us go now to sample estimates and investigate relations between $\beta(\tilde{\nu})$ and $\hat{\beta}(\tilde{\nu})$. We first need the following lemma:

Lemma 2

(i) $\|E_x\| = O_p\left(\frac{n}{\sqrt{T}}\right)$

(ii) $\|E_{xy}\| = O_p\left(\frac{\sqrt{n}}{\sqrt{T}}\right)$

Proof. We have:

$$\|E_x\|^2 \leq \text{trace}\left[E_x' E_x\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} e_{x,ij}^2$$

Taking expectations, we obtain:

$$\text{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} e_{x,ij}^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{E}\left[e_{x,ij}^2\right] \leq \frac{n^2 K}{T} = O\left(\frac{n^2}{T}\right)$$

We further have $\|E_{xy}\|^2 = \sum_{i=1}^{n} e_{xy,i}^2$. Taking expectations:

$$\text{E}\left[\sum_{i=1}^{n} e_{xy,i}^2\right] = \sum_{i=1}^{n} \text{E}\left[e_{xy,i}^2\right] \leq \frac{nK}{T} = O\left(\frac{n}{T}\right)$$

The results follow from the Markov's inequality. Q.E.D.

Lemma 3 Under assumptions A1-3 and CR1-2,

$$\|\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu})\| = O\left(\frac{\sqrt{n}}{\tilde{\nu}\sqrt{T}}\right) \text{ as } n, T \to \infty$$

Proof. From (7) we have

$$\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu}) = S_x(\tilde{\nu})^{-1} S_{xy} - \Sigma_x(\tilde{\nu})^{-1}\Sigma_{xy}$$

and hence also

$$\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu}) = S_x(\tilde{\nu})^{-1}[S_{xy} - \Sigma_{xy}] + S_x(\tilde{\nu})^{-1}\Sigma_{xy} - \Sigma_x(\tilde{\nu})^{-1}\Sigma_{xy}$$

Using again the identity (10), we get

$$\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu}) = S_x(\tilde{\nu})^{-1}[S_{xy} - \Sigma_{xy}] + S_x(\tilde{\nu})^{-1}[\Sigma_x(\tilde{\nu}) - S_x(\tilde{\nu})]\Sigma_x(\tilde{\nu})^{-1}\Sigma_{xy}$$

whence

$$\|\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu})\| \leq \|S_x(\tilde{\nu})^{-1}\| \left(\|S_{xy} - \Sigma_{xy}\| + \|\Sigma_x(\tilde{\nu}) - S_x(\tilde{\nu})\| \, \|\beta(\tilde{\nu})\|\right)$$

Using Lemma 2, the bound (8) and the fact that $\|S_x(\tilde{\nu})^{-1}\| \leq \frac{1}{\tilde{\nu}}$, we get the desired result. Q.E.D.

Summing up, since $\|X_t\| = O_p(\sqrt{n})$, Lemma 1 tells us that $\beta(\tilde{\nu})'X_t$ converges to the optimal projection $\beta'X_t$ if $\frac{\tilde{\nu}}{n} \to 0$ as $n, T \to \infty$. Lemma 3 tells us that $\hat{\beta}(\tilde{\nu})'X_t$ converges to $\beta(\tilde{\nu})'X_t$ if $\frac{\tilde{\nu}}{n}\sqrt{T} \to \infty$ as $n, T \to \infty$. If $\tilde{\nu}$ meets both conditions we hence obtain a consistent estimate from $\hat{\beta}(\tilde{\nu})'X_t$. The following lemma combines both estimates (using the triangular inequality):

Lemma 4 Under the assumptions A1-3 and CR1-2, if $\frac{\tilde{\nu}}{n} \to 0$ and $\frac{\tilde{\nu}}{n}\sqrt{T} \to \infty$ as $n, T \to \infty$, then:

$$\hat{\beta}(\tilde{\nu})'X_t = \beta'X_t + O_p\left(\frac{\tilde{\nu}}{n}\right) + O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right) \text{ as } n \to \infty,$$

A suitable choice for the regularization parameter is $\tilde{\nu} = \alpha \, nT^{-(\frac{1}{2}-\delta)}$, where α is a constant.

The Proposition 2 is now established using Proposition 1, Lemma 4 and the definition of $\tilde{\nu}$.

7 Appendix B

An alternative to matrix inversion for computing regression estimates is provided by iterative methods as, for example, the so-called *Landweber iteration* which was initially developed for solving the normal equations in (7).

To insure convergence the algorithm is applied to regressors with norm smaller than 1. Since our regressors are standardized, this is insured by using the rescaled regressors $\tilde{X} = \frac{1}{\sqrt{n(p+1)(T-h-p)}}X$, and hence estimate the corresponding regression coefficients $\tilde{\beta} = \sqrt{n(p+1)(T-h-p)}\beta$.

Starting from the normal equation of the ordinary least squares, we can rewrite it as $\tilde{\beta} = \tilde{\beta} + \tilde{X}'y - \tilde{X}'\tilde{X}\tilde{\beta}$ and try to solve it through the successive approximations scheme

$$\tilde{\beta}^{(j+1)} = \tilde{\beta}^{(j)} + \tilde{X}'y - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)}; \qquad j = 0, 1, \ldots \tag{11}$$

A nice feature of the Landweber iteration is that it can be easily extended to cope with additional constraints or penalties, and in particular those used in Ridge or Lasso regression. As concerns the Lasso functional (5), Daubechies, Defrise, and De Mol (2004) have recently the following *thresholded Landweber iteration*

$$\tilde{\beta}^{(j+1)} = \mathbf{S}_\nu(\tilde{\beta}^{(j)} + \tilde{X}'y - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)}); \qquad j = 0, 1, \ldots \tag{12}$$

where the thresholding operator is acting on a vector componentwise by performing the soft-thresholding operation defined by (6) and is thus given by

$$\mathbf{S}_\nu(\tilde{\beta}) = [S_\nu(\tilde{\beta}_i)]_{i=1,\cdots,n}; \qquad i = 1, \ldots, n \tag{13}$$

This operation enforces the sparsity of the regression coefficients in the sense that all coefficients below the threshold $\nu/2$ are set to zero. The scheme (12) has been proved in Daubechies, Defrise, and De Mol (2004) to converge to a minimizer of the Lasso functional (5). Let us remark that this functional is not strictly convex when the null-space of $\tilde{X}$ is not reduced to zero and therefore the minimizer of (5) is not necessarily unique.

8 Appendix C

Table A: Data Transformation

	Definition	Transformation
1	$X_{it} = Z_{it}$	no transformation
2	$X_{it} = \Delta Z_{it}$	monthly difference
4	$X_{it} = \ln Z_{it}$	log
5	$X_{it} = \Delta \ln Z_{it} \times 100$	monthly growth rate
6	$X_{it} = \Delta \ln \frac{Z_{it}}{Z_{it-12}} \times 100$	monthly difference of yearly growth rate

			Lasso Selection*	
Code	**Description**	Transf.	IP	CPI
a0m052	Personal income (AR, bil. chain 2000 $)	5		
A0M051	Personal income less transfer payments (AR, bil. chain 2000 $)	5		II
A0M224 R	Real Consumption (AC) A0m224gmdc	5		
A0M057	Manufacturing and trade sales (mil. Chain 1996 $)	5		
A0M059	Sales of retail stores (mil. Chain 2000 $)	5		
IPS10	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX	5		
IPS11	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL	5		
IPS299	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS	5		
IPS12	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS	5		
IPS13	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS	5		
IPS18	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS	5		
IPS25	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT	5		
IPS32	INDUSTRIAL PRODUCTION INDEX - MATERIALS	5	II	
IPS34	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS	5		II
IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS	5	II	
IPS43	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)	5		
IPS307	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES	5		
IPS306	INDUSTRIAL PRODUCTION INDEX - FUELS	5		
PMP	NAPM PRODUCTION INDEX (PERCENT)	1		
A0m082	Capacity Utilization (Mfg)	2		
LHEL	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)	2		
LHELX	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF	2		
LHEM	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)	5		
LHNAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)	5		
LHUR	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)	2		
LHU680	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)	2		
LHU5	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)	5		
LHU14	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)	5		
LHU15	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)	5		
LHU26	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)	5		
LHU27	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS,SA)	5		
A0M005	Average weekly initial claims, unemploy. insurance (thous.)	5		
CES002	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE	5		
CES003	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING	5		
CES006	EMPLOYEES ON NONFARM PAYROLLS - MINING	5		II
CES011	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION	5		
CES015	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING	5		
CES017	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS	5		
CES033	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS	5		
CES046	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING	5		
CES048	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES	5		
CES049	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE	5	II	
CES053	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE	5		
CES088	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES	5	I	I
CES140	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT	5		
A0M048	Employee hours in nonag. establishments (AR, bil. hours)	5		
CES151	AVG WEEKLY HOURS OF PROD. OR NONSUPERV. WORKERS ON PRIVATE NONFARM	1		
CES155	AVG WEEKLY HOURS OF PROD. OR NONSUPERV. WORKERS ON PRIVATE NONFARM	2		
aom001	Average weekly hours, mfg. (hours)	1		
PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)	1		I
HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA	4		
HSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.	4	II	
HSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.	4		II
HSSOU	HOUSING STARTS:SOUTH (THOUS.U.)S.A.	4		
HSWST	HOUSING STARTS:WEST (THOUS.U.)S.A.	4		I
HSBR	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)	4	I	
HSBNE	HOUSES AUTHORIZED BY BUILD. PERMITS:NORTHEAST(THOU.U.)S.A	4	II	I
HSBMW	HOUSES AUTHORIZED BY BUILD. PERMITS:MIDWEST(THOU.U.)S.A.	4		I-II
HSBSOU	HOUSES AUTHORIZED BY BUILD. PERMITS:SOUTH(THOU.U.)S.A.	4		I
HSBWST	HOUSES AUTHORIZED BY BUILD. PERMITS:WEST(THOU.U.)S.A.	4		
PMI	PURCHASING MANAGERS' INDEX (SA)	1		
PMNO	NAPM NEW ORDERS INDEX (PERCENT)	1	I	
PMDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)	1		
PMNV	NAPM INVENTORIES INDEX (PERCENT)	1		II
A0M008	Mfrs' new orders, consumer goods and materials (bil. chain 1982 $)	5		
A0M007	Mfrs' new orders, durable goods industries (bil. chain 2000 $)	5		
A0M027	Mfrs' new orders, nondefense capital goods (mil. chain 1982 $)	5		
A1M092	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 $)	5	I	
A0M070	Manufacturing and trade inventories (bil. chain 2000 $)	5	I	

Code	Description	Transf.	Lasso Selection* IP	CPI
A0M077	Ratio, mfg. and trade inventories to sales (based on chain 2000 $)	2		
FM1	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)	6		
FM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,GP&BD MMMFS&SAV&SM TIME DEP(BIL$,	6		I
FM3	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S&INST ONLY MMMFS)(BIL$,SA)	6		
FM2DQ	MONEY SUPPLY - M2 IN 1996 DOLLARS (BCI)	5	I	I
FMFBA	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)	6		
FMRRA	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)	6		
FMRNBA	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL$,SA)	6		
FCLNQ	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1996 DOLLARS (BCI)	6		
FCLBMC	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)	1		II
CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)	6		
A0M095	Ratio, consumer installment credit to personal income (pct.)	2		
FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)	5		
FSPIN	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)	5	II	
FSDXP	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)	2	I	
FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)	5		
FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)	2		
CP90	Cmmercial Paper Rate (AC)	2		
FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)	2		
FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)	2		
FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)	2		
FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)	2		
FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)	2		
FYAAAC	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)	2		
FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)	2		II
scp90	cp90-fyff	1		II
sfygm3	fygm3-fyff	1	I	
sFYGM6	fygm6-fyff	1		
sFYGT1	fygt1-fyff	1		
sFYGT5	fygt5-fyff	1		
sFYGT10	fygt10-fyff	1	II	
sFYAAAC	fyaaac-fyff	1		
sFYBAAC	fybaac-fyff	1		
EXRUS	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)	5		
EXRSW	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)	5		
EXRJAN	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)	5		
EXRUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)	5		
EXRCAN	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)	5		
PWFSA	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)	6		
PWFCSA	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)	6		
PWIMSA	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)	6		
PWCMSA	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)	6		
PSM99Q	INDEX OF SENSITIVE MATERIALS PRICES (1990=100)(BCI-99A)	6	I	
PMCP	NAPM COMMODITY PRICES INDEX (PERCENT)	1	II	II
PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)	6		I
PU83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)	6		
PU84	CPI-U: TRANSPORTATION (82-84=100,SA)	6	I	
PU85	CPI-U: MEDICAL CARE (82-84=100,SA)	6	II	
PUC	CPI-U: COMMODITIES (82-84=100,SA)	6		
PUCD	CPI-U: DURABLES (82-84=100,SA)	6		
PUS	CPI-U: SERVICES (82-84=100,SA)	6		
PUXF	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)	6		
PUXHS	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)	6		
PUXM	CPI-U: ALL ITEMS LESS MIDICAL CARE (82-84=100,SA)	6		
GMDC	PCE,IMPL PR DEFL:PCE (1987=100)	6		
GMDCD	PCE,IMPL PR DEFL:PCE; DURABLES (1987=100)	6		
GMDCN	PCE,IMPL PR DEFL:PCE; NONDURABLES (1996=100)	6		
GMDCS	PCE,IMPL PR DEFL:PCE; SERVICES (1987=100)	6		
CES275	AVG HOURLY EARNINGS OF PROD. OR NONSUPERV. WORKERS ON PRIVATE NONFARM	6		I
CES277	AVG HOURLY EARNINGS OF PROD. OR NONSUPERV WORKERS ON PRIVATE NONFARM	6	II	
CES278	AVG HOURLY EARNINGS OF PROD. OR NONSUPERV. WORKERS ON PRIVATE NONFARM	6		
HHSNTN	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)	2		

*We indicate when forecasting IP or CPI, the variable has been selected by Lasso regression at the beginning (I), 1970 : 1, and/or and the end (II), 2001 : 12, of the out-of-sample evaluation period.