

Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?

Christine De Mol^a, Domenico Giannone^{a,b,c}, Lucrezia Reichlin^{c,d}

^a ECARES, Université Libre de Bruxelles, Belgium

^b European Central Bank, Germany

^c CEPR, UK

^d London Business School, UK

article info

Article history:

Available online 28 August 2008

JEL classification:

C11

C13

C33

C53

Keywords:

Bayesian shrinkage

Bayesian VAR

Ridge regression

Lasso regression

Principal components

Large cross-sections

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 conditions for consistency of the forecast based on Bayesian regression as the cross-section and the sample size become large. This analysis serves as a guide to establish a criterion for setting the amount of shrinkage in a large cross-section.

© 2008 Elsevier B.V. All rights reserved.

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 et al., 2005; Giannone et al., 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, 2006, 2005a; Wright, 2003) but, surprisingly, few papers explore the performance of Bayesian regression in forecasting with high-dimensional data. Exceptions are Stock and Watson (2005a), who consider normal Bayes estimators for orthonormal regressors, and Giacomini and White (2006), who provide an empirical example in which a Bayesian regression with a large number of predictors is compared with principal component regression (PCR).

Bayesian methods are part of the traditional econometrician toolbox and offer a natural solution to overcoming 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 et al., 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 in the literature.

This paper is a first step towards filling this gap. We analyze Bayesian regression methods under Gaussian and double-exponential priors and study their forecasting performance on the standard “large” macroeconomic dataset that has been used to establish properties of principal-component-based forecasts (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 a Gaussian prior, the posterior mode solution is such that all variables in the panel are given non-zero coefficients. Regressors,

Corresponding address: European Central Bank, Directorate General Research, Kaiserstrasse 29, D-60311 Frankfurt am Main, Germany. Tel.: +49 69 1344 7849; fax: +49 69 1344 8553.

E-mail address: Domenico.Giannone@ecb.int (D. Giannone).

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 impose unit weight to the dominant ones and zero to the others. The double-exponential prior, 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 density to be either large or zero. As a result, it favors the recovery of 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 a Gaussian prior, it is easy to compute the maximizer of the posterior density. Under such a prior with independent and identically distributed (i.i.d.) regression coefficients, the solution amounts to solving a penalized least-squares problem with a penalty proportional to the sum of the squares of the coefficients, i.e. to a so-called Ridge regression problem. Under a double-exponential prior, however, there is no analytical form for the maximizer of the posterior density, but we can exploit the fact that, under such a prior with i.i.d. coefficients, the solution amounts to a Lasso regression problem, i.e. to penalized least-squares with a penalty proportional to the sum of the absolute values of the coefficients. Several algorithms have been proposed for Lasso regression. In our empirical study, we have used two algorithms, recently proposed, which work without limitations of dimensionality: LARS (Least Angle Regression) developed by Efron et al. (2004) and the Iterative Landweber scheme with soft-thresholding at each iteration developed by De Mol and Defrise (2002) and Daubechies et al. (2004).

An interesting feature of Lasso regression is that it combines variable selection and parameter estimation. The estimator depends in a nonlinear way 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 and requiring computationally demanding algorithms, such as the one proposed by Fernandez et al. (2001) in the context of Bayesian Model Averaging and applied by Stock and Watson (2005a) to macroeconomic forecasting with large cross-sections.

Although Gaussian and double-exponential Bayesian regressions rely on different estimation strategies, 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 which are highly correlated and are characterized by similar mean-square errors. Moreover, these forecasts are highly correlated with those produced by principal components, also with similar mean-square errors: they do well when PCR does well. Hence, although the Lasso prior leads to the selection of few variables, the forecasts obtained from these informative targeted predictors do not outperform PCR based on few principal components.

In order to understand these results, we study the asymptotic properties of the forecast based on Bayesian regression as the cross-section and the sample size become large. This double-asymptotic analysis has been applied by recent literature to the case of PCR (Bai, 2003; Bai and Ng, 2002; Forni et al., forthcoming, 2004; Stock and Watson, 2002a,b) but never to Bayesian regression. This analysis is however important to understanding the performance of this method for large panels and also as a guide to setting shrinkage parameters as the dimension of the panel changes. Here we will limit the analysis to the Bayesian regression based on a Gaussian prior and show that, under very general conditions, consistency is achieved provided that the degree of shrinkage increases with the cross-sectional dimension.

The conditions under which we show consistency require that most of the regressors are informative about the future of the variable to forecast. This condition is satisfied in the particular case in which the data follow an approximate factor structure, the case for which the literature has shown consistency for PCR. The approximate factor structure imposes a high degree of collinearity in the data that persists as we add series to the panel. Intuitively, under those assumptions, if the prior is chosen appropriately in relation with n , Bayesian regression under normality will give larger weight to the principal components associated with the dominant eigenvalues and therefore will produce results which are similar to PCR.

Our empirical work shows, moreover, that Lasso forecasts, although based on regression on few variables are as accurate and as highly correlated with PCR forecasts as are those obtained under normality. This result may seem puzzling, but it can be explained by the fact that our panel is highly collinear. Under collinearity, few variables, if selected appropriately, should capture the essence of covariation. In this case, we expect them to be strongly correlated with principal components and, as the latter, to span the space of the pervasive common factors. Under collinearity, however, we expect the selection not to be stable and to be very sensitive to minor perturbation of the data. In this sense, we do not expect variable selection to provide results which lead to clearer economic interpretation than principal components or Ridge regression.

The paper is organized as follows. Section 2 introduces the problem of forecasting using large cross sections. The Section 3 reports the results of the out-of-sample exercise for the three methods considered: principal components, Bayesian regression with normal and with double-exponential priors. The Section 4 reports asymptotic results for the Gaussian prior case. The Section 5 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 $Z_t \in \mathbb{R}^n$, z_{1t}, \dots, 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

$$y_{tChj} \in \text{proj}_{f_{tChj}} y_{tChj}$$

where $f_{tChj} \in \text{span}\{Z_{t,s}; s \in \{0, 1, 2, \dots, g\}\}$ is a potentially large information set at time t and $y_{tChj} \in \mathbb{R}^{h_{tChj}}$, $h_{tChj} \in \mathbb{N}$, $h_{tChj}/n \rightarrow 0$ as $n \rightarrow \infty$ is a filtered version of z_{it} , for a specific $i \in \{1, \dots, n\}$.

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:

$$y_{tChj} \in Z_{t,0} \beta_0 + \dots + Z_{t,p} \beta_p + u_{tChj}$$

where $\beta_0, \dots, \beta_p \in \mathbb{R}^{h_{tChj}}$ and $X_t \in \mathbb{R}^{(p+1) \times h_{tChj}}$. The implied forecast is given by $y_{tChj}^0 \in \mathbb{R}^{h_{tChj}}$ and the implied forecast error is $u_{tChj} \in \mathbb{R}^{h_{tChj}}$. The latter is assumed to be orthogonal to $z_{it,s}$ for $s \in \{0, 1, \dots, p\}$ and $i \in \{1, \dots, n\}$.

Given a sample of size T , we will denote by $X \in \mathbb{R}^{T \times (p+1) \times h_{tChj}}$ the matrix of observations for the predictors and by $y \in \mathbb{R}^{T \times h_{tChj}}$ the matrix of observations on the dependent variable. The regression coefficients are typically estimated by Ordinary Least Squares (OLS), $\hat{\beta}^{OLS} \in \mathbb{R}^{(p+1) \times h_{tChj}}$, and the forecast is given by $\hat{y}_{tChj}^{OLS} \in \mathbb{R}^{h_{tChj}}$. When the size of the information set, n , is large, such a projection involves the estimation of a large number of parameters. This implies a loss of degrees of freedom and a poor forecast (“curse of dimensionality problem”). Moreover, if the number of regressors is larger than the sample size, $n \times (p+1) > T$, OLS is not feasible.

To solve this problem, the literature proposes computing the forecast as a projection on the first few principal components (Forni et al., 2005; Giannone et al., 2004, 2008; Stock and Watson, 2002a,b).

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

$$S_X V = D V \quad (1)$$

where $D = \text{diag}.d_1, \dots, d_{n.p.C1}/$ is a diagonal matrix having on the diagonal the eigenvalues of $S_X = \frac{1}{T-h-p} X^0 X$ in decreasing order of magnitude and $V = [V_1, \dots, V_{n.p.C1}]$ is the $n.p.C1 \times n.p.C1$ matrix whose columns are the corresponding normalized eigenvectors.¹ The normalized principal components (PC) are defined as:

$$\hat{\theta}_{it} = \frac{1}{d_i} V_i^0 X_t \quad (2)$$

for $i = 1, \dots, N$ where N is the number of non-zero eigenvalues.²

If most of the interactions among the variables in the information set are due to a 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 captured by projecting on the dependent variable itself (or on a small set of predictors). In such situations, few principal components provide a good approximation of the underlying factors. The principal component forecast is defined as:

$$y_{tChjt}^{PC} = \text{proj}_{f_t} y_{tChjt} = \text{proj}_{f_t} y_{tChjt} \quad (3)$$

where $f_t = \text{span}\{\hat{\theta}_{1t}, \dots, \hat{\theta}_{rt}\}$, with $r = n.p.C1$, 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 Section 4). 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 we follow consists instead in imposing limits on the length of θ through priors and estimating the parameters as the posterior mode. The parameters are then used to compute the forecasts. Here we consider two alternatives: a Gaussian and a double-exponential prior.

Let us assume that $u_t \sim \text{i.i.d.}(\mathcal{N}(0, \frac{2}{u}), \frac{2}{u})$, with known variance $\frac{2}{u}$; then, under a Gaussian prior $\theta \sim \mathcal{N}(0, \frac{2}{u})$, and assuming for simplicity that all parameters are shrunk to zero, i.e. $\theta_0 = 0$, we have:

$$\theta^{bay} = X^0 X C \frac{2}{u} \frac{1}{\theta_0} X^0 y$$

The corresponding forecast is then computed as:

$$\hat{y}_{TCChjt}^{bay} = X_T^0 \theta^{bay}$$

¹ The eigenvalues and eigenvectors are typically computed on $\frac{1}{T-h-p} \sum_{t=DpC1}^T X_t X_t^0$ (see for example Stock and Watson (2002a)). We instead compute them on $\frac{1}{T-h-p} X^0 X = \frac{1}{T-h-p} \sum_{t=DpC1}^T X_t X_t^0$ for comparability with the other estimators considered in the paper.

² Note that $N = \min\{n.p.C1; T-h-p\}$.

In the case in which the parameters are i.i.d.,³ i.e. $\theta_0 = 0$, the estimates are equivalent to those produced by penalized Ridge regression with parameter $D = \frac{2}{u}$. Precisely⁴:

$$\theta^{bay} = \arg \min_k \|y - Xk\|_2^2 + C \|k\|_2^2$$

It is known that there exist close relationships between OLS, PCR, penalized and Bayesian regression – see e.g. the book by Hastie et al. (2001) for a more detailed discussion of the connections between the different methods. For example, if the prior belief on the regression coefficients is that they are i.i.d., the forecast can be represented as a weighted sum of the projections on the principal components:

$$X_T^0 \theta = \sum_{i=1}^N w_i \hat{\theta}_{it} \theta_i \quad (4)$$

where $\theta_i = \frac{1}{d_i} V_i^0 X^0 y = T-h-p$ is the OLS regression coefficient of y on the i th principal component. For OLS we have $w_i = 1$ for all i . For the Bayesian estimates $w_i = \frac{d_i}{d_i C \frac{2}{u} + 1}$, where $D = \frac{2}{u}$. For the PCR regression we have $w_i = 1$ for $i \leq r$, and zero otherwise.

OLS, PCR and Gaussian Bayesian regression give non-zero weight to all variables. An alternative is to select variables. For Bayesian regression, variable selection can be achieved by a double-exponential prior, which, in the case of a zero-mean i.i.d. prior, is equivalent to the method that is sometimes called Lasso regression (an acronym for “least absolute shrinkage and selection operator”).⁵ 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:

$$\theta^{lasso} = \arg \min_k \|y - Xk\|_2^2 + C \sum_{i=1}^N |k_i| \quad (5)$$

where $D = 1$, being the scale parameter of the prior density⁶ (see e.g. Tibshirani (1996) and Fu (1998)).

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 a known analytical form. In particular, let us consider the case in which the regressors are the principal components of X . In this case, the Lasso solution can be cast in the form (4) with $w_i \theta_i$ replaced by $S_i \theta_i$, where S_i is the soft-threshold defined by

$$S_i = \begin{cases} C - 2 & \text{if } |k_i| \geq 2 \\ 0 & \text{if } |k_i| < 2 \\ 2 & \text{if } |k_i| \geq 2 \end{cases} \quad (6)$$

³ 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 appropriate for comparison with principal components.

⁴ In what follows we will denote by $\|k\|_2$ the L^2 matrix norm, i.e. for every matrix A , $\|A\|_2 = \sqrt{\lambda_{\max}(A^0 A)}$ where $\lambda_{\max}(A^0 A)$ is the maximum eigenvalue of $A^0 A$. For vectors k , $\|k\|_2$ denotes the Euclidean norm.

⁵ It should be noted however that Lasso is actually the name of an algorithm proposed in Tibshirani (1996) for finding a minimizer of (5).

⁶ We recall here that the variance of the prior density is proportional to 2^{-2} .

Hence, this sparse solution is obtained by setting to zero all coefficients β_i which in absolute value lie below the threshold $=2$ 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 hard-thresholding function 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 just a few observables rather than on a few linear combinations of the observables. Note that with such non-Gaussian priors the model is not invariant under orthogonal linear transformation of the data.

Notice also that, unlike in Ridge and PC regressions, where the regressors are weighted independently of the choice of the series to be forecasted, in Lasso regression the selection and shrinkage depend on that choice.

Methods described by Eq. (4) will perform well provided that no truly relevant coefficients β_i are observed for $i > r$, because in principal component regression they will not be taken into account and in Ridge regression 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 β_i with $i > r$. Lasso regression solves this problem.

Unfortunately, in the general case, the mode of the posterior distribution has no analytical form and has to be computed using numerical methods such as the Lasso algorithm of Tibshirani (1996) or quadratic programming based on interior point methods as advocated in Chen et al. (2001). Two efficient alternatives to the Lasso algorithm, which work without limitations of dimensionality also for a sample size T smaller than the number of regressors n , $p \ll 1$, have been developed more recently by Efron et al. (2004) under the name LARS (Least Angle Regression)⁷ and by De Mol and Defrise (2002) and Daubechies et al. (2004) who use instead an Iterative Landweber scheme with soft-thresholding applied at each iteration step.⁸

In the next Section we study 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 dataset 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, for a total of $n = 131$ variables.⁹

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 et al. (2004, 2008).

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

$$z_{IP,tCh}^h \equiv \frac{IP_t - IP_{t-Ch}}{IP_{t-Ch}} \quad \text{and} \quad z_{IP,tC1} \equiv \frac{IP_t - IP_{t-12}}{IP_{t-12}}$$

$$z_{CPI,tCh}^h \equiv \frac{CPI_t - CPI_{t-Ch}}{CPI_{t-Ch}} \quad \text{and} \quad z_{CPI,tC1} \equiv \frac{CPI_t - CPI_{t-12}}{CPI_{t-12}}$$

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

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

$$\hat{IP}_{TChjT}^h \equiv \hat{\beta}_{IP,TChjT}^h \quad \text{and}$$

and

$$\hat{b}_{TChjT}^h \equiv \hat{\beta}_{CPI,TChjT}^h \quad \text{and} \quad \hat{IP}_T$$

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

$$MSFE^h \equiv \frac{1}{T_1 - T_0} \sum_{t=T_0+h}^{T_1} \frac{1}{h} \sum_{\tau=0}^{h-1} (\hat{IP}_{TChjT}^h - IP_{TCh}^h)^2$$

and

$$MSFE_{IP}^h \equiv \frac{1}{T_1 - T_0} \sum_{t=T_0+h}^{T_1} \frac{1}{h} \sum_{\tau=0}^{h-1} (\hat{IP}_{TChjT}^h - IP_{TCh}^h)^2$$

The sample has a monthly frequency and ranges from 1959:01 to 2003:12. The evaluation period is 1970:01 to 2002:12. $T_1 \equiv 2003/12$ is the last available point in time, $T_0 \equiv 1969/12$ and $h \equiv 12$. 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. For all methods we report results for $p \equiv 0$ (no lags of the regressor) which is the one typically considered in macroeconomic applications. Qualitative results are not affected by this choice.¹⁰

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 $r \in \{1; 3; 5; 10; 25; 50; 75\}$ principal components. The case $r \equiv 0$ is the forecast implied from a random walk with drift on the log of IP and the annual CPI inflation.

We report MSFE relative to the random walk, and the variance of the forecasts relative to the variance of the series of interest.¹¹ The MSFE is also reported for two sub-samples: the first half of the evaluation period 1970–1985, and the second half 1985–2002. These results help us understand the relative performance of the methods for the cases where the predictability of key macroeconomic time series has dramatically decreased (on this point, see D'Agostino et al. (2006)). Results are reported in Table 1.

⁷ The LARS algorithm has also been used in econometric forecasting by Bai and Ng (2008) who also use it for selecting variables to form principal components.

⁸ 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 et al. (2004).

⁹ A full description of our dataset is given in a separate appendix containing supplementary material about this paper and available on request or from the website <http://homepages.ulb.ac.be/~dgiannon/>.

¹⁰ Empirical results supporting this claim are reported in a separate appendix containing supplementary material about this paper and available on request or from the website <http://homepages.ulb.ac.be/~dgiannon/>.

¹¹ We limit the empirical evaluation to point forecasts as it is standard in the literature on principal components forecasts. The theoretical results derived in the next section are also limited to point forecasts.

Table 1
Principal component forecasts

Industrial production	Number of principal components						
	1	3	5	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 ^a	0.23	0.70	0.79	0.97	1.28	1.43	1.78
Consumer price index	Number of principal components						
	1	3	5	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 ^a	0.36	0.55	0.61	0.63	0.69	0.89	1.69

MSFE are relative to a the Naïve, Random Walk, forecast.

^a The variance of the forecast relative to the variance of the series.

Let us start with the entire 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 increases and can become even 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.

The empirical literature on principal component regression has also considered the inclusion of the past of the variable of interest to capture series specific dynamics. The inclusion of those additional regressors does not affect qualitative results and in particular does not significantly improve the accuracy of the forecasts.¹²

Let us now do a similar exercise for the i.i.d. Gaussian prior (Ridge regression). Note, that, for $h \geq 1$, this case corresponds to a row of a VAR of order one. The Gaussian prior works well for the case $p \geq 0$ considered here.¹³

For the Bayesian Gaussian (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 $1 - \alpha$ of the variance of the variable to be forecast. We report results for different values of α (the associated values of λ , which are kept fixed for the whole out-of-sample evaluation period, are also reported). Notice that $\alpha = 1$ corresponds to the random walk since, in this case, all coefficients are set to zero. The other extreme, α 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.

The Ridge forecast performs well for a range of α 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 those obtained with principal

component regression. Moreover, the forecasts produced by Ridge regressions are smoother than the PC forecasts, which is a desirable property.

The last line of the table shows the correlation among Ridge forecasts and principal component forecasts.¹⁴ 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 regression 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 regression uses all eigenvalues in decreasing importance instead of truncating after r as in the principal components case. Notice also that, for inflation, with α 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 λ , 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 $k \in \{1; 3; 5; 10; 25; 50; 75\}$ non-zero coefficients.¹⁵

Results, reported in Table 3, show that good forecasts are obtained with a limited number of predictors, between 5 and 25. As for Ridge forecast, 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 high correlation of the forecasts suggest that all three methods are capturing similar features of the data. In particular, 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.

The similarity between forecasts based on PC and Ridge can be explained by collinearity among predictors. In fact, since the covariance of our data is characterized by few dominant eigenvalues, PC and Ridge forecasts, 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 Section 4 on the basis of the asymptotic analysis.

The result for the Lasso forecast is less straightforward to interpret since it 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 suggests that our data are highly collinear. Under collinearity, a few variables, if appropriately selected, should capture the essence of the covariation of the data and, as principal components, span approximately the space of the pervasive common factors. However, under these circumstances, we should also expect the selection to be unstable and very sensitive to minor perturbations of the data. With collinear data structure, variable selection methods are unlikely to provide results that are more interpretable than principal components or Ridge regressions from the economic point of view.

¹² Results are available on request. Similar results have also been reported in D'Agostino and Giannone (2007).

¹³ Incidentally, for the case $p > 0$, let us observe that it might be useful to shrink more the coefficients of additional lagged regressors, as, for example, with the Minnesota prior (Doan et al., 1984; Litterman, 1986). An additional feature of the Litterman priors is to shrink less the coefficients associated with the variable to forecast. This can be helpful when series specific dynamics have significant forecasting power. The study of such more refined priors goes beyond the scope of the present empirical analysis which is meant as a first assessment of the general performance of the methods.

¹⁴ For the principal component forecasts we use $r \geq 10$. We obtain similar results also for $r \in \{3; 5\}$, i.e. when PC forecasts perform well.

¹⁵ An alternative, closer in spirit to the exercise with a Gaussian prior, is to select the prior λ at the beginning of the evaluation and then keep it fixed over the evaluation sample. This alternative strategy provides qualitatively similar results. See the appendix with supplementary material.

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
	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 ^a	0.71	0.63	0.57	0.49	0.39	0.29	0.19	0.12	0.07
Correlation with PC forecasts $r_D 10/$	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
	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 ^a	0.41	0.35	0.32	0.28	0.24	0.19	0.13	0.08	0.05
Correlation with PC forecasts $r_D 10/$	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.

^a The variance of the forecast relative to the variance of the series.

Table 3
Lasso forecasts

Industrial production							
	Number of non-zero coefficients						
	1	3	5	10	25	50	75
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 ^a	0.07	0.16	0.24	0.40	0.53	0.65	0.79
Correlation with PC forecasts $r_D 10/$	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 ^a	0.05	0.09	0.18	0.26	0.33	0.39	0.50
Correlation with PC forecasts $r_D 10/$	0.05	0.64	0.81	0.85	0.84	0.68	0.44

MSFE are relative to a the Naive, Random Walk, forecast.

^a The variance of the forecast relative to the variance of the series.

We examined the variables selected for $k = 10$ at the beginning and at the end of the out-of-sample evaluation period.¹⁶ Two main results emerge from this analysis. First, only some of the selected variables coincide with 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, although selected variables generally belong to the same economic category.

We have two conjectures about these results. The fact that variables are not clearly interpretable and that the procedure selects different variables at different points of the sample is, as mentioned above, the consequence of collinearity. The latter result

also suggests temporal instability. Notice, however, that temporal instability does not affect the relative performance of principal components and Ridge forecasts with respect to Lasso forecasts. This suggests that principal components and Ridge forecasts, 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 Bayesian regression and PCR are methods that help us solve the curse of dimensionality problem which typically arises when trying to extract relevant information from a large number of predictors.

For PCR, the literature has analyzed the asymptotic properties for the size of the cross-section n and the sample size T going to infinity under assumptions that essentially impose that, as we increase the number of time series, the sources of common dynamics remain limited (Bai, 2003; Bai and Ng, 2002; Forni et al., forthcoming, 2005; Stock and Watson, 2002a,b). Double

¹⁶ These variables are reported in the last two columns of the table describing the database contained in the appendix containing supplementary material about this paper and available on request or from the website <http://homepages.ulb.ac.be/~dgiannon/>.

asymptotics for Bayesian regression, on the other hand, has never been studied and this is a relevant analysis for understanding its behaviour when using a large number of predictors. In what follows, we will consider double (n, T) asymptotics for the case of the Gaussian prior, under conditions that are more general than those considered for PCR in the literature mentioned above. As we will see, our assumptions impose that the optimal forecast and the observable predictors depend on a finite number of unobserved factors.

For the sake of simplicity, we will assume throughout this Section that no lags of the regressors are used in the forecasting regression. In the notation of Section 2, this means that we set $p = 0$. In this case, X_t coincides with the predictors at time t , $Z_t = z_{1,t}, \dots, z_{n,t}$. All results, however, apply straightforwardly to the case in which lagged predictors are also included, i.e. p different from zero.

Let us first assume that the forecast of y_t depends on a finite number of unobserved factors.

Assumption A. $y_{tCh} = F_t' C y_{tCh}$, where V_{tCh} is orthogonal to X_t for all n and where the factors $F_t = f_{1,t}, \dots, f_{r,t}$ are a r -dimensional stationary process with covariance matrix $E F_t F_t' = I_r$.

Consider the forecast based on the projection on the unobserved factors F_t :

$$y_{tCh|t} = F_t'$$

Under **Assumption A**, the forecast $y_{tCh|t}$ is optimal in the sense that, due to the assumption of orthogonality between the residuals V_{tCh} and the observed predictors X_t , its accuracy cannot be improved using the information available at time t . For fixed n , the optimal forecast is unfeasible, even with infinite sample size T , since the factors are unobserved. We assume that the observed predictors are related to the common factors as follows:

Assumption B. X_t has the following representation:

$$X_t = F_t' C_t$$

where

- (i) the residuals v_t are a n -dimensional stationary process with covariance matrix $E v_t v_t' = I_n$ of full rank for all n ;
- (ii) the matrix loading the factors is a non-random matrix of dimension $n \times r$ and of full rank r for each n ;
- (iii) the residuals v_t are orthogonal to the factors F_t .

In **Assumption B** the predictors are decomposed into two parts. One part (F_t) is driven by the factors which are informative about the future of the target variable. The residuals (v_t) can be considered as the component of the predictors that is not informative. For convenience we assume that the two components are orthogonal. The assumption that the non-informative residuals are of full rank entails that there are no redundant predictors. This ensures that when we increase the number of predictors we do not duplicate information.

Under **Assumptions A** and **B**, we have $X_t = E X_t X_t' / D = I_n$ and $y_{tCh} = E y_{tCh} y_{tCh}' / D = I_n$. Because of **Assumption B(i)**, X_t is invertible for all n . Consequently, for a given number n of predictors, the population OLS regression coefficient $\beta = X_t^{-1} y_{tCh}$ is unique and the forecast is given by:

$$y_{tCh|t} = X_t^{-1} y_{tCh} = \beta$$

Let us first derive conditions on the shrinkage parameter that will allow us to obtain consistent forecasts from Bayesian regression under Gaussian priors. We will need the additional **Assumption C** that ensures that the elements of the sample covariances of X_t with itself and with y_t converge uniformly to their population counterpart; see **Appendix A** for details.

Let us consider the prediction based on the Gaussian prior, u_t i.i.d.: $N(0, \sigma_u^2)$ and $\sigma_u^2 \rightarrow 0$. We have the following result:

Proposition 1. Under **Assumptions A–C**, and if $\liminf_{n,T \rightarrow \infty} \frac{\min_{k \leq n} \sigma_k^2}{k} > 0$, we have for $n, T \rightarrow \infty$:

$$X_t^{-1} y_{tCh|t} = y_{tCh|t} + O_p \left(\frac{p_{1,n}}{n} \sqrt{\frac{p_{2,n}}{n}} \right) \quad \text{as } n, T \rightarrow \infty$$

where $p_{1,n} = \max_{k \leq n} \sigma_k^2$ and $p_{2,n} = \min_{k \leq n} \sigma_k^2$.

Proof. See **Appendix A**.

Proposition 1 indicates that the behavior of the Bayesian forecast under a Gaussian prior is governed by the quantities $p_{1,n}$ and $p_{2,n}$, which in turn are related to the information content of the observable predictors X_t with respect to the factors F_t . If the factors are pervasive throughout the predictors' cross-section with non decreasing weights, then $p_{1,n}$ and $p_{2,n}$ go to infinity with n . We assume that they increase linearly with n :

Assumption D.

$$0 < \liminf_{n \rightarrow \infty} \frac{1}{n} \min_{k \leq n} \sigma_k^2 < \limsup_{n \rightarrow \infty} \frac{1}{n} \max_{k \leq n} \sigma_k^2 < \infty$$

Under **Assumption D** all the predictors are informative for the factors F_t and hence they all help improve the forecast accuracy. In this case the sample forecast converges to its population counterpart. Precisely:

Corollary 1. Under the assumptions of **Proposition 1**, and if **Assumption D** holds, then:

$$X_t^{-1} y_{tCh|t} = y_{tCh|t} + O_p \left(\frac{1}{n} \right) \quad \text{as } n, T \rightarrow \infty$$

To achieve consistency, a suitable choice for the prior is $\sigma_u^2 = c n T^{-\frac{1}{2}}$, with $0 < c < 2$ and c an arbitrary constant. Under this condition on the prior we have

$$X_t^{-1} y_{tCh|t} = y_{tCh|t} + O_p(1/n) \quad \text{as } n, T \rightarrow \infty$$

where $n T \rightarrow \infty$ and $0 < c < 2$. Let us stress here that no restriction on the relative path of divergence of T and n is needed in order to achieve consistency. In this sense the estimates are viable also when the size of the cross-section n is much larger than the sample size T .

Corollary 1 tells us that, under the factor structure assumption, the Bayesian regression should use a prior that shrinks increasingly all regression coefficients to zero as the number of predictors rises. This is because, if the factors are pervasive, 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 condition $\liminf_{n,T \rightarrow \infty} \frac{\min_{k \leq n} \sigma_k^2}{k} > 0$ requires that all the regression coefficients should be shrunk at the same asymptotic rate.

In the empirical exercise, the condition $\liminf_{n,T \rightarrow \infty} \frac{\min_{k \leq n} \sigma_k^2}{k} > 0$ is satisfied since we used the i.i.d. prior ($\sigma_u^2 = c n T^{-\frac{1}{2}}$). Moreover, from **Corollary 1**, consistency requires that the shrinkage parameter $\sigma_u^2 = c n T^{-\frac{1}{2}}$ grows asymptotically at a rate equal to the number of predictors n . Although this is an asymptotic condition that is difficult to assess empirically on the basis of a finite cross-section and sample size, the empirical results appear to confirm that roughly,

¹⁷ This is a generalization to the dynamic case of the assumptions defining an approximate factor structure given by Chamberlain and Rothschild (1983). Bai (2003), Bai and Ng (2002) and Stock and Watson (2002a) give similar conditions.

with the dominant eigenvalues is not distorted asymptotically whereas the effect of the smallest ones goes to zero asymptotically.

Notice that the rates of consistency of the Bayesian forecasts are slower than the ones derived for principal components by Bai (2003) under the assumption that the non-informative component ε_t is idiosyncratic. The reason is that, as we have seen, the assumptions required to achieve consistency of the forecast based on Bayesian regression are more general than those implied by an approximate factor structure. In particular, the convergence of the Bayesian to the optimal forecast is achieved in the case in which Assumption E holds for $0 < \lambda_{\min} \leq 1$. This can be viewed as a sort of “weak factor structure” since $\lambda_{\min} \rightarrow 0$ can be unbounded as $n \rightarrow \infty$. More interestingly, convergence to the population forecast based on the observed predictors y_{tChj} holds under arbitrary correlation structure among the non-informative component ε_t . Such generality gives flexibility to the method but at the price of a slower rate of convergence.

This result suggests that the properties of alternative methods suitable for forecasting with a large number of predictors can be studied under more general conditions than those used in the recent literature on principal components.¹⁸

Let us remark in concluding this section that we have only studied theoretical properties of point forecasts. Under the assumption that the data follow an approximate factor structure, prediction intervals for principal components regressions are derived in Bai and Ng (2006). For the Bayesian regression, predictive intervals can be computed from the posterior distribution, although theoretical properties for large cross-sections are not known.

5. Conclusions and open questions

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

We have considered the Gaussian and the double-exponential prior and showed 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 et al., 1984; Litterman, 1986) and related literature. We explore this conjecture in a related paper (Banbura et al., forthcoming).

In the asymptotic analysis, we have considered the Gaussian prior case. For that case, we have shown $\lambda_{\min} \rightarrow 0$ 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 amount of shrinkage for the double-exponential case, on the other hand, has been exclusively empirical. It is designed 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. Selected variables, however, are not clearly interpretable, typically not the ones that a macroeconomist would include in a VAR. Moreover, the selected variables change over time. These results suggest that our data, which correspond to the typical macroeconomic dataset used for macroeconomic policy analysis, is characterized by collinearity. Under collinearity

we should expect both that few appropriately selected variables capture the bulk of the covariation and that the selection is sensitive to minor perturbations of the data. In these circumstances we should not expect to obtain results that are more interpretable, from the economic point of view, than principal components or Ridge regression, but we should expect comparable forecasting performance. To explore in more depth these conjectures, we should extend the double-asymptotic analysis that we have provided for the Gaussian case to the double-exponential Bayesian regression. We intend to do this in further work.

Acknowledgments

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 an anonymous referee, Marta Banbura, Michel Defrise, James Hamilton, James Naeson, 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. Replication files as well as an appendix containing supplementary material are available at: <http://homepages.ulb.ac.be/~dgiannoni/> or <http://homepages.ulb.ac.be/~leicli/>.

Appendix A. Proof of Proposition 1

Throughout this Section we will maintain the simplifying assumption $p \rightarrow 0$ introduced in Section 4. All results still hold for the case where p lags are included, by simply replacing n by $n \cdot p \rightarrow 0$.

Denote:

- by y_t the generic variable to be forecast as $y_t \in \mathbb{R}^1$;
- the covariance matrix of the regressors as $\Sigma_x = E(X_t X_t')$. The sample equivalent will be denoted by $S_x = X'X/T$ and the estimation error by $E_x = y - X\beta$. These matrices are of dimension $n \times n$;
- the covariance matrix of the regressors and the variable to be predicted by $\Sigma_{xy} = E(X_t y_t)$. The sample equivalent will be denoted by $S_{xy} = X'y/T$ and the estimation error by $E_{xy} = y - X\beta$. These matrices are of dimension $n \times 1$.

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

Assumption C. There exists a finite constant K , such that for all $T \geq N$ and $i, j \geq N$

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

as $T \rightarrow \infty$, where $e_{x,ij}$ denotes the i, j th entry of E_x and $e_{xy,i}$ denotes the i th entry of E_{xy} . Sufficient conditions can be found in Forni et al. (forthcoming).

We consider here only the case of an i.i.d. Gaussian prior on the coefficients and we denote by $\lambda \rightarrow 0$ the rescaled penalization in the Ridge regression.

¹⁸ Within the framework of principal component regression, this line of research has been pursued by Onatski (2006) who studies the properties of principal components when factors are weak.

Remark 1. Notice that this does not imply that we lose in generality. Indeed, in the case of a non-i.i.d. prior, we can always redefine the regression in terms of $\tilde{X}_t = \frac{1}{\sqrt{k_0 k}} X_t$. Then the corresponding rescaled regression coefficients, $\beta = \frac{1}{\sqrt{k_0 k}} \beta_0$, will be i.i.d. with prior variance $k_0 k$. Moreover, under [Assumption B](#), the transformed regressors \tilde{X}_t have the representation

$$\tilde{X}_t = \beta_0 F_t + \epsilon_t$$

where $\beta_0 = \frac{1}{\sqrt{k_0 k}}$ and $\epsilon_t = \frac{1}{\sqrt{k_0 k}} \epsilon_t$. The assumption $\liminf_{n,T \rightarrow \infty} \frac{\min_{i,j} \epsilon_{xy,ij}^2}{k_0 k} > 0$ ensures that the transformed regressors still satisfy [Assumptions D](#) and [E](#) when the original regressors do.

Defining $\beta_x = \beta_0 / D$ and the sample equivalent $S_x = \beta_0 / D S_x$, we are interested in the properties of β_x and β_0 which are solutions of the following linear system of equations:

$$\begin{aligned} \beta_x \cdot \beta_0 / D &= \beta_{xy} \\ S_x \cdot \beta_0 / D &= S_{xy} \end{aligned} \quad (9)$$

Notice that β_0 / D is the population OLS regression coefficient and β_0 / D its sample counterpart. For $\beta_0 > 0$, we have the Ridge regression coefficients.

Lemma 1. Under the assumptions of [Proposition 1](#) we have

$$k \cdot \beta_0 / k = k \cdot \beta_0 / D \cdot D \cdot \frac{1}{2n} \quad (10)$$

and

$$k \cdot \beta_0 / k \leq \frac{1}{2n} \quad \text{as } n \rightarrow \infty \quad (11)$$

Proof. We have

$$k \cdot \beta_0 / k \leq \frac{1}{2n} \quad \text{as } n \rightarrow \infty$$

First notice that for any vector v of the form $v = \frac{1}{\sqrt{k_0 k}} w$, i.e. orthogonal to the null-space of β_0 , we have the inequality $\min_{i,j} \epsilon_{xy,ij}^2 / k \leq k \cdot \beta_0 / k$ since β_0 is positive definite. Then taking $w = \frac{1}{\sqrt{k_0 k}} v$, the inequality becomes

$$\min_{i,j} \epsilon_{xy,ij}^2 / k \leq k \cdot \beta_0 / k \quad (12)$$

and holds for any w in the range of β_0 , i.e. of the form $w = \beta_0 v$. Now, replacing w in (12) by $\beta_0 v$, we get, for any v

$$k \cdot \beta_0 / k \leq \frac{1}{2n} \quad \text{as } n \rightarrow \infty$$

where $\frac{1}{2n} = \frac{\max_{i,j} \epsilon_{xy,ij}^2}{k_0 k}$ and $\frac{1}{2n} = \frac{\min_{i,j} \epsilon_{xy,ij}^2}{k_0 k}$. Since $k \cdot \beta_0 / k \leq \frac{1}{2n}$, we get (10). The inequality $k \cdot \beta_0 / k \leq \frac{1}{2n}$ is straightforward. Now,

$$\begin{aligned} \beta_0 / D &= \frac{1}{D} \cdot \frac{1}{\sqrt{k_0 k}} \cdot \frac{1}{\sqrt{k_0 k}} \cdot \frac{1}{\sqrt{k_0 k}} \\ &= \frac{1}{D} \cdot \frac{1}{\sqrt{k_0 k}} \cdot \frac{1}{\sqrt{k_0 k}} \cdot \frac{1}{\sqrt{k_0 k}} \end{aligned}$$

thanks to the matrix identity

$$A^{-1} B^{-1} D A^{-1} B = A/B^{-1} \quad (13)$$

Hence

$$k \cdot \beta_0 / k \leq \frac{1}{2n} \quad \text{as } n \rightarrow \infty$$

Replacing w in (12) by $\beta_0 v$ and using the bound (10), we easily obtain the bound (11).

In contrast to what happens for $\beta_0 = 0$, notice that the Ridge parameter β_0 introduces a bias which tends to zero for large cross-sectional dimensions provided that the asymptotic behavior of β_0 as $n \rightarrow \infty$ is appropriately tuned.

Let us now consider the sample estimates and investigate the relationship between β_0 / D and β_0 / D . We first need the following lemma:

Lemma 2. (i) $k E_x k \leq O_p \left(\frac{1}{\sqrt{n}} \right)$

$$(ii) k E_{xy} k \leq O_p \left(\frac{1}{\sqrt{n}} \right)$$

Proof. We have:

$$k E_x k^2 = \text{trace} \left(E_x^0 E_x \right) = \sum_{i=1}^p \sum_{j=1}^p e_{x,ij}^2$$

Taking expectations, we obtain:

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

We also have $k E_{xy} k^2 \leq \sum_{i=1}^p \sum_{j=1}^p e_{xy,ij}^2$. Taking expectations, we get:

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

The results follow from the Markov inequality.

Lemma 3. Under the [Assumptions A–C](#), we have

$$k \cdot \beta_0 / k \leq \frac{1}{2n} \quad \text{as } n \rightarrow \infty$$

Proof. From (9) we have

$$\beta_0 / D = \beta_0 / D S_x \cdot \beta_0 / D^{-1} S_{xy} = \beta_x \cdot \beta_0 / D^{-1} S_{xy}$$

and hence also

$$\beta_0 / D = \beta_0 / D S_x \cdot \beta_0 / D^{-1} S_{xy} = \beta_x \cdot \beta_0 / D^{-1} S_{xy}$$

Using again the identity (13), we get

$$\beta_0 / D = \beta_0 / D S_x \cdot \beta_0 / D^{-1} S_{xy} = \beta_x \cdot \beta_0 / D^{-1} S_{xy}$$

whence

$$k \cdot \beta_0 / k \leq \frac{1}{2n} \quad \text{as } n \rightarrow \infty$$

Using [Lemma 2](#), the bound (10) and the fact that $k S_x \cdot \beta_0 / D^{-1} k = 1$, we get the desired result.

We can now combine the results of [Lemma 1](#) and [Lemma 3](#), by means of the triangular inequality, and use the fact that $k X_t k \leq O_p \left(\frac{1}{\sqrt{n}} \right)$ to establish the following lemma, which is a simple corollary of the previous ones.

Lemma 4. Under [Assumptions A–C](#), we have, as $n, T \rightarrow \infty$,

$$\begin{aligned} X_t^0 \cdot \beta_0 / D &= X_t^0 \cdot \beta_0 / D \cdot \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \\ &= \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \end{aligned}$$

[Proposition 1](#) is now established using [Lemma 4](#) and the definition of $\beta_0 = \frac{1}{\sqrt{k_0 k}}$.

By a proper choice of the regularization parameter λ as a function of n and T , we will have to ensure that both terms tend to zero as $n/T \rightarrow 1$. This is done in Corollary 1, under a supplementary assumption on the asymptotic behaviour of $\lambda_{1,n}$ and $\lambda_{2,n}$ (Assumption D).

Appendix B

An alternative to matrix inversion for computing regression estimates is provided by iterative methods as, for example, the so-called *Landweber iteration* scheme, which can be modified to cope with the penalties used in Ridge and Lasso regression.

To ensure convergence of this algorithm the norm of the sample matrices X must be smaller than 1. Since our regressors are standardized, this condition is fulfilled when using the rescaled regressors $\tilde{X} = \frac{1}{\sqrt{n \cdot p \cdot C(1/T) \cdot h(p)}}$, and hence estimating the corresponding regression coefficients as $\tilde{\beta} = \frac{1}{\sqrt{n \cdot p \cdot C(1/T) \cdot h(p)}}$.

Starting from the normal equation of the ordinary least squares, we can rewrite it as $\tilde{\beta} = \tilde{X}'\tilde{X}\tilde{\beta} = \tilde{X}'\tilde{y}$ and, starting from arbitrary $\tilde{\beta}^{(0)}$, try to solve it through the successive approximations scheme

$$\tilde{\beta}^{(j+1)} = \tilde{X}'\tilde{X}\tilde{\beta}^{(j)} + \tilde{X}'\tilde{y} - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)} \quad j = 0, 1, \dots \quad (14)$$

which is the standard Landweber iteration. A nice feature of this scheme is that it can be easily extended to cope with additional constraints or penalties, and in particular with those used in Ridge or Lasso regression. As concerns the Lasso cost function (5), Daubechies et al. (2004) have recently proposed the following *thresholded Landweber iteration*

$$\tilde{\beta}^{(j+1)} = \tilde{X}'\tilde{X}\tilde{\beta}^{(j)} + \tilde{X}'\tilde{y} - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)} \quad j = 0, 1, \dots \quad (15)$$

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

$$\tilde{\beta}^{(j+1)} = \tilde{X}'\tilde{X}\tilde{\beta}^{(j)} + \tilde{X}'\tilde{y} - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)} \quad j = 0, 1, \dots \quad (16)$$

This operation enforces the sparsity of the regression coefficients in the sense that all coefficients below the threshold $\lambda = 2$ are set to zero. The scheme (15) has been proved in Daubechies et al. (2004) to converge to a minimizer of the Lasso cost function (5). Let us remark that this functional fails to be strictly convex when the null-space of \tilde{X} is not reduced to zero, in which case the minimizer of (5) is not necessarily unique.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jeconom.2008.08.011.

References

- Bai, J., 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71 (1), 135–171.
 Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. *Econometrica* 70 (1), 191–221.

- Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74 (4), 1133–1150.
 Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146 (2), 304–317.
 Banbura, M., Giannone, D., Reichlin, L., 2007. Bayesian VARs with Large Panels, Discussion Paper 6326, Center for Economic Policy Research. *Journal of Applied Econometrics* (forthcoming).
 Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305–1324.
 Chen, S.S., Donoho, D., Saunders, M., 2001. Atomic decomposition by basis pursuit. *SIAM Review* 43, 129–159.
 D'Agostino, A., Giannone, D., 2007. Comparing alternative predictors based on large-panel factor models, Discussion Paper 6564, Center for Economic Policy Research.
 D'Agostino, A., Giannone, D., Surico, P., 2006. (Un)Predictability and Macroeconomic Stability, Working Paper Series 605, European Central Bank.
 Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* 57, 1416–1457.
 De Mol, C., Defrise, M., 2002. A note on wavelet-based inversion methods. In: Nashed, M.Z., Scherzer, O. (Eds.), *Inverse Problems, Image Analysis and Medical Imaging*. American Mathematical Society, pp. 85–96.
 Doan, T., Litterman, R., Sims, C.A., 1984. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews* 3, 1–100.
 Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. *Annals of Statistics* 32, 407–499.
 Fernandez, C., Ley, E., Steel, M.F.J., 2001. Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100 (2), 381–427.
 Forni, M., Giannone, D., Lippi, M., Reichlin, L., 2007. Opening the black box: Structural factor models with large cross-sections, Working Paper Series 712, European Central Bank. *Econometric Theory* (forthcoming).
 Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* 82, 540–554.
 Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2004. The generalized dynamic factor model consistency and rates. *Journal of Econometrics* 119 (2), 231–255.
 Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association* 100, 830–840.
 Fu, W.J., 1998. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7, 497–516.
 Giacomini, R., White, H., 2006. Tests of conditional predictive ability. *Econometrica* 74 (6), 1545–1578.
 Giannone, D., Reichlin, L., Sala, L., 2004. Monetary policy in real time. In: Gertler, M., Rogoff, K. (Eds.), *NBER Macroeconomics Annual*. MIT Press, pp. 161–200.
 Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics* 55 (4), 665–676.
 Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Springer-Verlag.
 Koop, G., Potter, S., 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.
 Onatski, A., 2006. Asymptotic distribution of the principal components estimator of large factor models when factors are relatively weak, Manuscript, Columbia University.
 Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 147–162.
 Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20, 147–162.
 Stock, J.H., Watson, M.W., 2005a. An Empirical Comparison Of Methods For Forecasting Using Many Predictors, Manuscript, Princeton University.
 Stock, J.H., Watson, M.W., 2005b. Implications of dynamic factor models for VAR analysis, NBER Working Papers 11467, National Bureau of Economic Research, Inc..
 Stock, J.H., Watson, M.W., 2006. Forecasting with Many Predictors. In: *Handbook of Economic Forecasting*, vol. 1. Elsevier, pp. 515–554 (chapter 10).
 Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series 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 (US).