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

This version: March 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 components 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 an approximate factor structure to establish a criterion for setting parameters in a large cross-section.

JEL Classification: C11,C13,C33,C53

Keywords: Bayesian VAR, ridge regressions, 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 for useful comments. 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

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., 2003; Giannone et al., 2004; Stock and Watson, 2002a,b). Surprisingly, the literature has not considered, amongst these methods, Bayesian regression.

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, Bayesian VAR have been advocated as a device for forecasting macroeconomic data (Doan et al., 1984; Litterman, 1986). These methods, however, have been applied to relatively small systems.

This paper analyzes Bayesian regression methods under Gaussian and double exponential prior and studies their forecasting performance on the standard "large" macroeconomic dataset that has been used to establish properties of principal components based forecast (Stock and Watson, 2002a,b).

Our two choices for the prior correspond to two interesting cases. The Gaussian density prior generates coefficients' posteriors implying that all variables in the panel are given some weight. This is the same as principal components regression, but while the Gaussian prior gives decreasing weight to the ordered eigenvalues of the covariance matrix of the data, principal components give positive weight to the dominant ones and zero to the others. The double exponential, on the other hand, puts more mass near zero and in the tails and this induces a tendency of the coefficients' posteriors to be either large or zero. As a result, one favors the recovery of a few large coefficients instead of many fairly small ones. Moreover, the double-exponential prior favors truly zero values instead of small ones, i. e. it favors *sparse* regression coefficients. This case is interesting because it results in variable selection rather than in variable aggregation as in the other two cases and, in principle, should be 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 a number of algorithms. Lasso regression is a method that combines variable selection and parameters estimation. The estimator is nonlinear and data-adaptive which may have advantages in some empirical situations.

Results based on an out-of-sample exercise show that, for a given grid of parameters, the forecasts of the two Bayesian methods considered are strongly correlated with that based on principal components and that mean squared errors are roughly similar.

To understand this result one should consider that the Bayesian regression with Gaussian i.i.d. prior on the coefficients is a ridge regression and the latter has a strict relation with principal components regression. Ridge and Lasso are

penalized regressions and, as principal components, they can be understood as regularization methods when the problem of computing OLS estimates is ill-conditioned by the presence of a cluster of eigenvalues of the covariance matrix close to zero (near-collinearity).

Collinearity is a typical feature of large panels of data. This feature, in the recent literature on large panels, is captured by assuming that, as the size of the cross-section n increases, few eigenvalues increase with it while the others are bounded. These assumptions define the approximate factor structure first introduced by Chamberlain and Rothschild (1983) and generalized by Forni and Lippi (2001) and Forni et al. (2000). Related assumptions have been introduced by Bai and Ng (2002) and Bai (2003). These authors have used these assumptions to derive the n and T asymptotic properties of the principal components forecast. It is then a natural starting point for our analysis to base the asymptotic of Bayesian regression in large panels on the same set of assumptions.

Under those assumptions, and for the Gaussian prior case, we derive conditions on the prior under which the forecast converges to the efficient one (i.e. the forecast under knowledge of the true parameters) as n, T go to infinity. The intuition of the result is that, with appropriately chosen priors, the Bayesian regression tends to give more weight to the dominant common factors. The asymptotic result gives guidance for the setting of the prior in a large cross-section framework.

Lasso forecasts are computed using the iterative algorithm proposed by Daubechies et al. (2004) which is feasible for the large dimensional cases (also $n >> T$). Forecast accuracy comparable with principal components is achieved by a regression on few variables. All forecasts – principal components, ridge and Lasso – are very correlated and this is intriguing since, as mentioned, the first two methods give positive weight to all variables while Lasso only weights some. We interpret the 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 analyze the n, T asymptotic properties of the Lasso forecast under the approximate factor structure and analyze the empirical potential of this method for economic applications.

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 prior and with double exponential. The fourth Section reports asymptotic results for the 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

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

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

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:

$$y_{t+h} = Z'_t\beta_0 + ... + Z'_{t-p}\beta_p + u_{t+h} = X'_t\beta + u_{t+h}$$

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 coefficient are typically estimated by Ordinary Least Square (OLS), $\hat{\beta}^{LS} = (X'X)^{-1}X'y$, and the forecasts 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 are 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 et al., 2003; Giannone et al., 2004, 2005; Stock and Watson, 2002a,b).

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

$$S_xV = VD \tag{1}$$

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[1]. The normalized principal components (PC) are defined as:

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

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

for $i = 1, \cdots, N$ where N is the number of non zero eigenvalues[2].

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 $r << np$, 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:

$$y_{t+h|t} = \text{proj}\left\{y_{t+h}|\Omega_t\right\} \approx \text{proj}\left\{y_{t+h}|\Omega_t^F\right\} \tag{3}$$

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 components forecast to the efficient forecast under assumptions defining an approximate factor structure (see next Section). Under those assumptions, once common factors are estimated via principal components, the projection implied by the forecast equation is computed by OLS treating the estimated factors as if they were observables.

The Bayesian approach consists instead in shrinking parameters through priors and in using the posterior of the parameters 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 shrank to zero, $\beta_0 = 0$, we have:

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

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, $\Phi_0 = \sigma_\beta^2 I$, this is equivalent to penalized, ridge regression with pa-

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

rameter $\nu = \frac{\sigma_u^2}{\sigma_\beta^2}$[3]. Precisely[4]:

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

If the priors on the regression coefficients are i.i.d. there is a very strict relation between Bayesian regression and principal components. OLS, principal component regression and Bayesian regressions can be represented as a weighted sum over the projection 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 $w = 1$ for all i. For the Bayesian estimates $w_i = \frac{d_i}{d_i + \frac{\nu}{T-h-p}}$, where $\nu = \frac{\sigma_u^2}{\sigma_\beta^2}$. For the principal components, we have $w_i = 1$ if $i \leq r$, and zero otherwise.

Linear methods select the relevant (stable) components of the solution and discard the others only on the basis of the signal-to-noise ratio and of the spectral properties of the matrix S_x to be inverted. The selection is thus performed independently of the specific observed data vector y (linear methods are not *data-adaptive*). A look at (4) shows that such methods will perform well provided that no truly significant coefficients α_i are observed for $i > r$, because those will not be taken into account. Bad performances is 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 $i > r$.

This drawback of linear methods can be overcome when resorting to nonlinear data-adaptive methods. A simple instance of a thresholding method can be derived by using non-gaussian priors on the regression coefficients.

A well known case is the 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) although Lasso is actually the name of an algorithm for finding the maximizer of the posterior proposed in Tibshirani (1996). In this particular i.i.d. prior case the method can also be seen as a penalized regression with a penalty in the coefficients involving the L_1 norm instead of the L_2 norm. Precisely:

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

[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 natural for allowing comparison with principal components.

[4]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.

where $\nu = \frac{1}{\tau}$ where τ is the scale parameter of the prior density (see e.g. Tibshirani, 1996). We recall here that the variance of the prior density is equal to $2\tau^2$.

Compared with the Gaussian density, the double exponential puts more mass near zero and in the tails and this induces a tendency of the coefficients' posteriors to be 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 posterior).

Unfortunately, the maximizer of the posterior distribution has no analytical form and has to be computed using numerical methods. An efficient alternative to the Lasso algorithm has been developed more recently by Efron et al. (2004) under the name LARS (Least Angle Regression) whereas quadratic programming based on interior point methods are advocated in Chen et al. (2001). We will use instead an Iterative Landweber scheme with soft thresholding at each iteration proposed in De Mol and Defrise (2002) and further analyzed in Daubechies et al. (2004) which works without limitations of dimensionality also for sample size T smaller than the number of regressors $n(p+1)$ (see Appendix B).

To gain intuition about Lasso regression, let us consider, as an example, the case of orthogonal regressors where the posterior 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, with not orthogonal regressors, the 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 non-Gaussian method is not invariant to orthogonal linear transformation of the data.

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 (2005), with the exception of a few series. 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 1.

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 in first differences of annual inflation following Giannone et al. (2004, 2005).

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

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

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 through as:

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

and

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

The predictions accuracy is evaluated using the mean squared 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 consumption price index (CPI).

Let us start from principal components. We report results for the choice of $r = 1, 3, 6, 10, 25, 50, 75$ 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.

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 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 D'Agostino et al. (2005). Results are reported in Table 1.

Table 1: Principal components forecasts

Consumer Price Index							
	Number of Principal Componennts						
	1	3	6	10	25	50	75
MFSE 1971-2002	0.89	0.64	0.54	0.54	0.66	1.10	1.54
MFSE 1971-1984	0.86	0.48	0.36	0.35	0.43	0.81	1.23
MFSE 1985-2002	0.99	1.12	1.10	1.12	1.33	1.96	2.50
Variance*	0.21	0.60	0.65	0.76	0.92	1.06	1.25

Consumer Price Index							
	Number of Principal Componennts						
	1	3	6	10	25	50	75
MFSE 1971-2002	0.61	0.56	0.59	0.69	0.83	1.09	1.47
MFSE 1971-1984	0.52	0.42	0.45	0.50	0.64	0.90	1.18
MFSE 1985-2002	1.01	1.22	1.29	1.56	1.69	2.02	2.81
Variance*	0.31	0.45	0.53	0.53	0.49	0.62	0.97

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 lagged regressors, as, for example, with the Minnesota prior (Doan et al., 1984; Litterman, 1986). This is beyond the scope of this empirical analysis here which is meant as a first assessment of the general performance of the methods.

For the 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 are also reported). Notice that $\kappa = 1$ corresponds to the random walk since, in this case, all regressors 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	5	23	60	140	290	580	1150	2350	6000
MFSE 1971-2002	0.99	0.71	0.60	0.56	0.56	0.58	0.64	0.72	0.84
MFSE 1971-1984	0.75	0.49	0.41	0.38	0.39	0.44	0.52	0.63	0.78
MFSE 1985-2002	1.70	1.37	1.20	1.10	1.04	1.01	1.00	0.99	0.99
Variance*	0.74	0.64	0.58	0.49	0.39	0.29	0.19	0.12	0.07

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	14	60	140	290	530	960	1750	3550	9250
MFSE 1971-2002	0.86	0.72	0.66	0.63	0.62	0.63	0.66	0.73	0.84
MFSE 1971-1984	0.70	0.57	0.52	0.51	0.51	0.54	0.59	0.68	0.82
MFSE 1985-2002	1.62	1.41	1.30	1.20	1.12	1.05	0.99	0.96	0.96
Variance*	0.41	0.35	0.32	0.28	0.24	0.19	0.13	0.08	0.05

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 components regression. Moreover, the forecasts

produced by ridge regressions are smoother than the PC forecasts, which is a desirable property.

As for the two sub-samples, results are also qualitatively similar to PC 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 PC case. This is not surprising since ridge uses all eigenvalues in decreasing importance instead of truncating after r as in the PC case. Notice also that, for inflation, for intermediate range of ν, 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 deliver a given number (k) of non zero coefficients in the initial sample $1960 - 1970$. We then use the same prior for the whole exercise. We look at $k \approx 5, 10, 25, 50, 60$ non zero regressors case.

Results, reported in Table 3 show that good forecasts are obtained with a limited number of predictors, between 5 and 10.

Table 3: Lasso forecasts

Industrial Production					
	Avg. Number of regressors				
	5	10	25	50	60
MSFE 1971-2002	0.62	0.60	0.67	0.71	0.80
MSFE 1971-1984	0.49	0.43	0.45	0.48	0.56
MSFE 1985-2002	1.02	1.10	1.32	1.40	1.52
Variance*	0.28	0.41	0.61	0.71	0.76

Consumer Price Index					
	Avg. Number of regressors				
	5	10	25	50	60
MSFE 1971-2002	0.66	0.61	0.69	0.80	0.89
MSFE 1971-1984	0.58	0.50	0.52	0.61	0.70
MSFE 1985-2002	1.04	1.15	1.47	1.68	1.75
Variance*	0.15	0.25	0.35	0.39	0.46

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

Finally, we look at the correlation among different forecasts. Results are reported for $r = 10$ principal components and 10 predictors for the first 10 years of observations for Lasso. For the Gaussian Bayesian model we select $\nu = 290$ which is the best performing both for IP and CPI. From Table 3 we can see the forecasts are highly correlated.

Comparable MSE 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

Table 4: Correlation among forecasts

Industrial Production					Consumer Price Index			
	PCA	ridge	Lasso			PCA	ridge	Lasso
PCA	1.00	–	–		PCA	1.00	–	–
Ridge	0.95	1.00	–		Ridge	0.95	1.00	–
Lasso	0.88	0.93	1.00		Lasso	0.87	0.93	1.00

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 the space of the pervasive common factors.

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 both CPI inflation and IP forecasts.

We have two conjectures about these results. The fact that variables are not clearly interpretable probably indicates that the panel contain clusters of correlated variables and the procedure select 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 point of the sample, implies temporal instability, but results imply that the latter does not affect the relative performance of methods such as ridge and principal components. This suggests that these methods, 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 principal components can be seen as regularization methods for near collinear data structures. Large panels of macroeconomic time series are typically highly collinear (Giannone et al., 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.

We suppose that X_t has the following representation[5]:

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

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

We will assume

$$y_{t+h} = \gamma F_t + u_{t+h}$$

where u_{t+h} is unforecastable and orthogonal to F_t. Hence

$$y_{t+h|t} = \gamma F_t$$

Following Forni et al. (2000, 2003, 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[6]. These conditions are a generalization to the dynamic case of the conditions defining an approximate factor structure given by Chamberlain and Rothschild (1983). Precisely:

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

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

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.

Under these conditions, it has been shown that (Forni et al., 2003, 2005):

$$X_T' \hat{\beta}^{bay} - \gamma F_T = o_p(1) \text{ as } n, T \to \infty,$$

where $\hat{F}_t$ are the PC estimates of the common factors and $\hat{\gamma}$ is estimated by OLS of y_{t+h} on $\hat{F}_t$. Bai (2003) and Stock and Watson (2002a) have shown the same result but under slightly different assumptions.

We study now the properties of the Bayesian estimates if the data are generated from an approximated factor structure. Let us first notice that under our assumptions:

$$y_{t+h} = X_t' \beta + u_{t+h}$$

[5]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.

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

Under the assumptions CR1 and CR2 we have $\beta_i \sim \frac{1}{n}$ and $\|\beta\| \sim \frac{1}{\sqrt{n}}$ for n large. In fact the population regression coefficients are given by

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

where $\Sigma_x = \mathrm{E}(X_t X_t')$ and $\Sigma_{xy} = \mathrm{E}(X_t y_{t+h})$. Under assumptions CR1-2 we have $\|(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\| \sim \frac{1}{n}$ while $\|\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda + I)^{-1}\| \sim \frac{1}{\sqrt{n}}$. This implies that the magnitude of the regression coefficients decreases with the number of regressors. The reason is that, if the factors are pervasive, then all variables are informative for the common factors and we should give weight to all of them. The result suggests that, under the factor structure assumption, the Bayesian regression should use a prior that, as the cross-section dimension increases, shrinks increasingly more regression coefficients to zero. The Proposition below gives conditions for the shrinkage parameter that allow to obtain consistent forecasts. We will need the additional Assumption A1 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 Under assumptions A1, CR1 and CR2, if $\liminf_{n\to\infty} \frac{\lambda_{min}(\Phi_0)}{\|\Phi_0\|} > 0$ then:

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

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

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.

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 et al. (2005) and, using a different set of assumptions by Bai (2003), and probably can be improved by imposing further assumptions.

The intuition of this result is very simple. The factor structure implies that there are few r dominant eigenvalues that diverge faster then the remaining smaller ones as the cross-section dimension increases. The parameter's prior

chosen as above insures that the effect of the factors associated with the dominant eigenvalues is not distorted asymptotically while for the smaller ones it is set to zero asymptotically. Clearly, as mentioned in the empirical Section, if there are few dominant eigenvalues, both Bayesian regression under Gaussianity and principal components regression will only give weight to the dominant eigenvalues.

Recall that in the case studied here, Bayesian regression with Gaussian prior is equivalent to ridge regression.

5 Conclusions and open questions

This paper has analysed the properties of Bayesian regression in large panels of time series and compared them to principal components regression.

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 PC regression and implies similar mean squared 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.

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 result 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. 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 analyzed 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.

References

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Gari Chamberlain and Michael Rothschild. Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica*, 51:1305–1324, 1983.

Scott Shaobing Chen, David Donoho, and Michael Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001.

Antonello D'Agostino, Domenico Giannone, and Paolo Surico. (un)predictability and macroeconomic stability. Macroeconomics 0510024, Economics Working Paper Archive EconWPA, October 2005.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1416–1457, 2004.

Christine De Mol and Michel Defrise. A note on wavelet-based inversion methods. In M. Z. Nashed and O. Scherzer, editors, *Inverse Problems, Image Analysis and Medical Imaging*, pages 85–96. American Mathematical Society, 2002.

Thomas Doan, Robert Litterman, and Christopher A. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–100, 1984.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32:407–499, 2004.

Mario Forni, Domenico Giannone, Marco Lippi, and Lucrezia Reichlin. Opening the black box: Structural factor models with large cross-sections. Manuscript, (www.dynfactors.org), 2005.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics*, 82:540–554, 2000.

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic factor model: one-sided estimtion and forecasting. CEPR Working Paper 9829, 2003. *Journal of the American Statistical Association*, forthcoming.

Mario Forni and Marco Lippi. The generalized dynamic factor model: representation theory. *Econometric Theory*, 17:1113–1141, 2001.

Domenico Giannone, Lucrezia Reichlin, and Luca Sala. Monetary policy in real time. In Mark Gertler and Kenneth Rogoff, editors, *NBER Macroeconomics Annual*, pages 161–200. MIT Press, 2004.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting gdp and inflation: The real time informational content of macroeconomic data releases. Technical Report 5178, CEPR Discussion Papers, August 2005.

Robert Litterman. Forecasting with bayesian vector autoregressions – five years of experience. *Journal of Business and Economic Statistics*, 4:25–38, 1986.

James. H. Stock and Mark. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:147–162, 2002a.

James. H. Stock and Mark. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economics Statistics*, 20:147–162, 2002b.

James H. Stock and Mark W. Watson. Implications of dynamic factor models for var analysis. Unpublished manuscript, Princeton University, 2005.

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

6 Appendix A: proof of Proposition 1

Denote by y_t the generic variable to be forecast, i.e. $y_t = z_{it}^h$

- the covariance matrix of the regressors as $\Sigma_x = \mathrm{E}(X_t X_t')$. The sample equivalent will be denoted by $S_x = X'X/(T-h-p)$. The estimation error will be denote by $E_x = S_x - \Sigma_x$. These matrices are of dimension $n(p+1) \times n(p+1)$.
- 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 $S_{xy} = X'y/(T-h-p)$. The estimation error will be denote by $E_{xy} = S_{xy} - \Sigma_{xy}$. These matrices are of dimension $n(p+1) \times 1$.

Assumption A1: 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 et al. (2005).

We can consider without loss of generality the case of iid prior on the coefficients and denote by $\tilde{\nu} = \frac{\sigma_u^2}{(T-h-p)\|\Phi_0\|}$ the rescaled associated penalization in

the ridge regression. In fact, in the case of non-iid prior, we can redefine the regression in terms of $\tilde{X}_t = \frac{1}{\sqrt{\|\Phi_0\|}}\Phi_0^{1/2}X_t$. The associated regression coefficients, $\tilde{\beta} = \sqrt{\|\Phi_0\|}\Phi_0^{-1/2}\beta$ have are now i.i.d. prior with variance $\|\Phi_0\|$. In additions the transformed regressors $\tilde{X}_t$ have the factor representation

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

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.

Define, $\Sigma_x(\tilde{\nu}) = \Sigma_x + \tilde{\nu}I_n$ 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:

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

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

Lemma 1 Under assumptions CR1-2 we have $\|\beta(\tilde{\nu})\| = O\left(\frac{1}{\sqrt{n}}\right)$ and

$$\beta(\tilde{\nu})'X_t = \gamma F_t + O_p\left(\frac{\tilde{\nu}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) \text{ as } n \to \infty,$$

Proof. We have:

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

hence

$$\|\beta(\tilde{\nu})\| \leq \left(\frac{\sqrt{\lambda_{max}(\Lambda'\Lambda)}}{\lambda_{min}(\Psi) + \tilde{\nu}}\right)\left(\frac{\|\Psi\| + \tilde{\nu}}{\lambda_{min}(\Lambda'\Lambda)}\right)\|\gamma\| = O\left(\frac{1}{\sqrt{n}}\right)$$

this proves the first result. Turning to the second statement, we have

$$\begin{array}{rl} \beta(\tilde{\nu})'X_t = & \gamma(\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\Lambda + I_r)^{-1}\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\Lambda F_t \\ & +\gamma(\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\Lambda + I_r)^{-1}\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\xi_t \end{array}$$

First notice that

$$(\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\Lambda + I_r)^{-1}\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\Lambda = I_r - \left(\Lambda'(\Psi + \tilde{\nu}I_n)^{-1}\Lambda + I_r\right)^{-1}$$

consequently

$$\gamma(\Lambda'(\Psi+\tilde{\nu}I_n)^{-1}\Lambda+I_r)^{-1}\Lambda'(\Psi+\tilde{\nu}I_n)^{-1}\Lambda F_t = \gamma F_t - \left(\Lambda'(\Psi+\tilde{\nu}I_n)^{-1}\Lambda + I_r\right)^{-1}\gamma F_t$$

where $\| \left(\Lambda'(\Psi+\tilde{\nu}I_r)^{-1}\Lambda + I_r\right)^{-1} \| \leq \frac{\|\Psi\|+\tilde{\nu}}{\lambda_{min}(\Lambda'\Lambda)} = O\left(\frac{\tilde{\nu}}{n}\right)$ as $n \to \infty$.

Turning to the second term :

$$\mathrm{E}\|\beta(\tilde{\nu})'\xi_t\|^2 = \beta(\tilde{\nu})'\Psi\beta(\tilde{\nu}) \leq \|\beta(\tilde{\nu})\|^2\|\Psi\| = O\left(\frac{1}{n}\right) \text{ as } n \to \infty$$

The desired result follows. Q.E.D.

For $\tilde{\nu} = 0$, we have that the optimal regression coefficient β provides consistent forecasts. The regularization parameter $\tilde{\nu}$ introduces a bias which tend to zero for large cross-sectional dimensions provided is does not increases faster than the cross-sectional dimension n. Let us move now to sample estimates and investigate relations between $\beta(\tilde{\nu})$ and $\hat{\beta}(\tilde{\nu})$. We first need the following lemma:

Lemma 2 Under Assumption A1, we have:

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

(i) $\|E_{xy}\| = O_p\left(\frac{\sqrt{n}}{\sqrt{T}}\right)$, as $n, T \to \infty$

Proof. We have:

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

Taking expectations, we obtain:

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

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

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

Results follow from the Markov inequality. Q.E.D.

We can now prove the main result.

Lemma 3 Under assumptions A1 and CR1-2, if $\frac{n}{\tilde{\nu}\sqrt{T}} = o(1)$ as $n, T \to \infty$, then

$$\|\hat{\beta}(\tilde{\nu})\| \leq \|\beta(\tilde{\nu})\| \left(1 + O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right)\right)$$

and

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

Proof. From Lemma 1 and the CR1-2 assumptions, we have

$$\|\Sigma_{xy}\|^{-1} = \|\Lambda\gamma'\|^{-1} \leq \sqrt{\lambda_{min}(\Lambda'\Lambda)^{-1}/\|\gamma\|^2} = O\left(\frac{1}{\sqrt{n}}\right)$$

Moreover,

$$\|\Sigma_x(\tilde{\nu})\|^{-1} \leq \frac{1}{\lambda_{min}(\Lambda'\Lambda)} = O\left(\frac{1}{n}\right) \text{ while } \|\Sigma_x^{-1}(\tilde{\nu})\| \leq \frac{1}{\tilde{\nu}}$$

Using Lemma 2, this implies that:

$$\frac{\|E_{xy}\|}{\|\Sigma_{xy}\|} = O_p\left(\frac{1}{\sqrt{T}}\right), \frac{\|E_x\|}{\|\Sigma_x(\tilde{\nu})\|} = O_p\left(\frac{1}{\sqrt{T}}\right), \text{ and } \|\Sigma_x(\tilde{\nu})^{-1}\|\|E_x\| = O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right)$$

For the normal equations, we obtain:

$$\left(I_n + \Sigma_x(\tilde{\nu})^{-1}E_x\right)\hat{\beta}(\tilde{\nu}) = \beta(\tilde{\nu}) + \Sigma_x(\tilde{\nu})^{-1}E_{xy}$$

Hence,

$$\hat{\beta}(\tilde{\nu}) = \left(I_n + \Sigma_x(\tilde{\nu})^{-1}E_x\right)^{-1}\left(\beta(\tilde{\nu}) + \Sigma_x(\tilde{\nu})^{-1}E_{xy}\right)$$

If $\frac{n}{\tilde{\nu}\sqrt{T}} = o(1)$, then for n, T large, we have $\|\Sigma_x(\tilde{\nu})^{-1}\|\|E_x\| < 1$. Consequently,

$$\begin{aligned}
\|\hat{\beta}(\tilde{\nu})\| &\leq \|\left(I_n + \Sigma_x(\tilde{\nu})^{-1}E_x\right)^{-1}\| \left(\|\beta(\tilde{\nu})\| + \|\Sigma_x(\tilde{\nu})^{-1}E_{xy}\|\right) \\
&\leq \tfrac{1}{1+o_p(1)}\left(\|\beta(\tilde{\nu})\| + \|\Sigma_x(\tilde{\nu})^{-1}E_{xy}\|\right) \\
&\leq \tfrac{1}{1+o_p(1)}\left(\|\beta(\tilde{\nu})\| + \|\Sigma_x(\tilde{\nu})^{-1}\|\|E_{xy}\|\right) \\
&\leq \tfrac{1}{1+o_p(1)}\left(\|\beta(\tilde{\nu})\| + O_p\left(\tfrac{1}{\sqrt{T}}\right)\|\Sigma_x(\tilde{\nu})^{-1}\|\|\Sigma_{xy}\|\right)
\end{aligned}$$

since $\|\Sigma_{xy}\| = \|\Sigma_x(\tilde{\nu})\beta(\tilde{\nu})\| \leq \|\Sigma_x(\tilde{\nu})\|\|\beta(\tilde{\nu})\|$, we get:

$$\|\hat{\beta}(\tilde{\nu})\| = \|\beta(\tilde{\nu})\| \left(1 + O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right)\right)$$

this proves the first statement. Turning to the second statement, we have:

$$\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu}) = \Sigma_x(\tilde{\nu})^{-1}E_{xy} - \Sigma_x(\tilde{\nu})^{-1}E_x\hat{\beta}(\tilde{\nu})$$

hence

$$\begin{aligned}
\|\hat{\beta}(\tilde{\nu}) - \beta(\tilde{\nu})\| &\leq \|E_{xy}\|\|\Sigma_{xy}\|^{-1}\|\Sigma_x(\tilde{\nu})^{-1}\|\|\Sigma_{xy}\| + \|\Sigma_x(\tilde{\nu})^{-1}E_x\|\|\hat{\beta}(\tilde{\nu})\| \\
&= O_p\left(\tfrac{1}{\sqrt{T}}\right)\|\Sigma_x(\tilde{\nu})^{-1}\|\|\Sigma_x(\tilde{\nu})\|\|\Sigma_x(\tilde{\nu})\|^{-1}\|\Sigma_{xy}\| + \|\Sigma_x(\tilde{\nu})^{-1}E_x\|\|\hat{\beta}(\tilde{\nu})\| \\
&= O_p\left(\tfrac{1}{\sqrt{T}}\right)O\left(\tfrac{n}{\tilde{\nu}}\right)\|\beta(\tilde{\nu})\| + O_p\left(\tfrac{n}{\tilde{\nu}\sqrt{T}}\right)\|\beta(\tilde{\nu})\|\left(1 + O_p\left(\tfrac{n}{\tilde{\nu}\sqrt{T}}\right)\right)
\end{aligned}$$

the desired result follows. Q.E.D.

Corollary 1 Under the assumptions of Lemma 3, we have:

$$\hat{y}_t := \hat{\beta}(\tilde{\nu})'X_t = \beta(\tilde{\nu})'X_t + O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right)$$

Proof. We have

$$\|(\beta(\tilde{\nu}) - \hat{\beta}(\tilde{\nu}))'X_t\| \leq \|X_t\|\|\beta(\tilde{\nu}) - \hat{\beta}(\tilde{\nu})\| = O_p\left(\sqrt{n}\right)\|\beta(\tilde{\nu})\|O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right)$$

the result follows from the fact that $\|\beta(\tilde{\nu})\| = O(\frac{1}{\sqrt{n}})$, see Lemma 1. Q.E.D.

Summing up, Lemma 1 tells us that $\beta(\tilde{\nu})X_t$ converges to the optimal projection 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 have that we obtain a consistent estimate from $\hat{\beta}(\tilde{\nu})'X_t$. Precisely:

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

$$\hat{\beta}(\tilde{\nu})'X_t = \gamma F_t + O_p\left(\frac{\tilde{\nu}}{n}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{n}{\tilde{\nu}\sqrt{T}}\right) \text{ as } n \to \infty,$$

A suitable choice for the regularization parameter is $\tilde{\nu} = \alpha n T^{-(\frac{1}{2}-\delta)}$, in this case we have:

$$\Delta_{nT}\left(\hat{\beta}(\tilde{\nu})' X_t - \gamma F_t\right) = O_p\left(1\right) \text{ as } n \to \infty,$$

where $\Delta_{nT} = \min\left\{\sqrt{n}, T^{\delta}, T^{(\frac{1}{2}-\delta)}\right\}$. Q.E.D

The desired result of Proposition 1 follows from the definition of $\tilde{v}$.

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 $\tilde{X} = \frac{1}{\sqrt{nT}} X$, and hence estimate the corresponding regression coefficients $\tilde{\beta} = \sqrt{nT}\beta$.

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

$$\tilde{\beta}^{(j+1)} = \tilde{\beta}^{(j)} + \tilde{X}'y - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)}; \qquad j = 0, 1, \ldots \tag{8}$$

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) , it is been shown recently in (Daubechies et al., 2004) that the following *thresholded Landweber iteration*

$$\beta^{(j+1)} = \mathbf{S}_{\nu}(\tilde{\beta}^{(j)} + \tilde{X}'y - \tilde{X}'\tilde{X}\tilde{\beta}^{(j)}); \qquad j = 0, 1, \ldots \tag{9}$$

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

$$\mathbf{S}_{\nu}(\tilde{\beta}) = S_{\nu}(\tilde{\beta}_i); \qquad i = 1, \ldots, n \tag{10}$$

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 (9) has been proved in Daubechies et al. (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
A0M051	Personal income less transfer payments (AR, bil. chain 2000 $)	5		
A0M224R	Real Consumption (AC) A0m224/gmdc	5		
A0M057	Manufacturing and trade sales (mil. Chain 1996 $)	5		
A0M059*	Sales of retail stores (mil. Chain 2000 $)	5		II
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		
IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS	5		
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		I
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		II
A0M048	Employee hours in nonag. establishments (AR, bil. hours)	5		
CES151	AVG WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS	1		
CES155	AVG WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS	2		
aom001	Average weekly hours, mfg. (hours)	1		
PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)	1		
HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA	4		
HSNE*	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.	4		I
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	
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		
PMDEL*	NAPM VENDOR DELIVERIES INDEX (PERCENT)	1	I	
PMNV*	NAPM INVENTORIES INDEX (PERCENT)	1	II	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		

Code	Description	Transf.	Lasso Selection* IP	CPI
A1M092*	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 $)	5	I-II	II
A0M070*	Manufacturing and trade inventories (bil. chain 2000 $)	5	I	
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$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,	6	I	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	II	II
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		I
FCLBMC*	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)	1		I
CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)	6		
A0M095	Ratio, consumer installment credit to personal income (pct.)	2		
FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)	5		
FSPIN*	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)	5	II	
FSDXP*	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)	2	I	
FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)	5		
FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)	2		
CP90	Cmmercial Paper Rate (AC)	2		
FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)	2		
FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)	2		
FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)	2		
FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)	2		
FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)	2		
FYAAAC*	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)	2		II
FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)	2		
scp90	cp90-fyff	1		
sfygm3*	fygm3-fyff	1	I	
sFYGM6	fygm6-fyff	1		
sFYGT1	fygt1-fyff	1		
sFYGT5	fygt5-fyff	1		
sFYGT10*	fygt10-fyff	1	II	
sFYAAAC	fyaaac-fyff	1		
sFYBAAC	fybaac-fyff	1		
EXRUS	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)	5		
EXRSW	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)	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	
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	II	
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		I
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 PRODUCTION OR NONSUPERVISORY WORKERS	6		I
CES277	AVG HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS	6		
CES278	AVG HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS	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.