Australian School of Business

Working Paper

Never Stand Still Australian School of Business

Australian School of Business Research Paper No. 2013 ECON 04

Forecasting using a large number of predictors: Bayesian model averaging versus principal components regression

Rachida Ouysse

www.asb.unsw.edu.au

Last updated: 15/04/13 CRICOS Code: 00098G

Forecasting using a large number of predictors: Bayesian model averaging versus principal components regression

Rachida Ouysse[*]

March 15, 2013

Abstract

We study the performance of Bayesian model averaging as a forecasting method for a large panel of time series and compare its performance to principal components regression (PCR). We show empirically that these forecasts are highly correlated implying similar mean-square forecast errors. Applied to forecasting Industrial production and inflation in the United States, we find that the set of variables deemed *informative* changes over time which suggest temporal instability due to collinearity and to the of Bayesian variable selection method to minor perturbations of the data. In terms of mean-squared forecast error, principal components based forecasts have a slight marginal advantage over BMA. However, this marginal edge of PCR in the average global out-of-sample performance hides important changes in the local forecasting power of the two approaches. An analysis of the Theil index indicates that the loss of performance of PCR is due mainly to its exuberant biases in matching the mean of the two series especially the inflation series. BMA forecasts series matches the first and second moments of the GDP and inflation series very well with practically zero biases and very low volatility. The fluctuation statistic that measures the relative local performance shows that BMA performed consistently better than PCR and the naive benchmark (random walk) over the period prior to 1985. Thereafter, the performance of both BMA and PCR was relatively modest compared to the naive benchmark.

1 Introduction

To overcome the challenges of dimensionality, many forecast approaches proceed by somehow reducing the number of predictors and three strands of the literature emerge. The first uses factor models and principal components regression (PCR). The second performs some sorts of variable selection to choose the "relevant" predictors and shrink to zero the coefficients of the noninformative predictors. Such methods include among others shrinkage regression such as *ridge* and *lasso.* The third is based on model averaging techniques and combines forecasts from all models.

[*] School of Economics, The University of New South Wales, Sydney 2052 Australia. Email: rouysse@unsw.edu.au.

Factor models are successfully used in forecasting with large number of predictors (Stock and Watson (2002a,b)). The diffusion index forecasts uses principal component regression (PCR) to summarize the information in all the predictors in a small number of factors (indexes). Forecasts that are based on these factors use information from all the predictors. However, this approach may not be optimal since the factors are created with no reference to the variables to be predicted. The factors are ranked according to the size of eigenvalues which is related to the amount of information extracted from the explanatory variables (not the target variable for forecasting) and it is possible that some factors associated with large eigenvalues have no explanatory power while some with small eigenvalues do have explanatory power for the dependent variable. To address this caveat, Bai and Ng (2008) use "targeted diffusion index forecasts" which target the estimation of the factor structure to the objective of forecasting by first pre-selecting a set of target predictors on which factor analysis is performed. They find that targeting predictors provides flexibility to adapt to parameter instability in the data and thus performs better than standard diffusion index forecasts. However, this approach still suffers from dimensionality if the factors are to be considered non-sequentially in the forecasting equation.

The idea of combining forecasts goes back to the work of Bates and Granger (1969) which pioneered the developing of theory of forecasts combination. See Clemen and Winkler (1986), Diebold and Lopez (1996), Hendry and Clements (2002) for excellent bibliography and reviews. Model averaging provides a kind of insurance against selecting a very poor model and can also avoid model selection instability by weighting/smoothing forecasts across several models, instead of relying entirely on a single model selected arbitrarily or by some model selection criterion. It thus enable considering different possible relationships between the predicted and the predictor variable. The analysis of the distribution of model averaging estimators can improve inference and prediction intervals and improves forecasts accuracy. However, there is no consensus on how to choose the combining weights. Many forecast combination methodologies use decision theoretic approaches to estimate the forecasts weights. These approaches can be classified into frequentist model averaging (FMA) and Bayesian model averaging. A few non-Bayesian methods for model averaging have been proposed in the literature. Hjort and Claeskens (2003) introduced a general class of FMA estimators which allow to perform valid classical inference in a model averaging context and define a framework for comparison with BMA estimators. In general the FMA weights are based on trade off between bias and variance and thus are based on model selection information criteria. Some examples of FMA estimators include Mallows model averaging of Hansen (2007) with weights based on the Mallows' criterion that minimizes mean-squared error (MSE) over the set of feasible forecast combinations, the smoothed information criteria estimator with either Akaike and BIC weights. Post model averaging and model selection inference has been studied in the literature, for example, Danilov and Magnus (2004) and Leeb and P`otcher (2006) and the asymptotic inference developed in Hjort and Claeskens (2003).

Bayesian Model averaging approach offers an alternative for exact finite sample inference by implicitly incorporating both model and parameters uncertainty into the distributions of the parameters. The weights applied in averaging the models are simply the posterior model weights. The statistical literature on BMA is enormous. Some examples include Raftery et al. (1997), Brown et al. (1999), Fernandez et al.

(2001) and George and Foster (2000). However, there have been relatively few papers in econometrics which adopt Bayesian model averaging to forecasting macroeconomic activity. Exceptions include studies that are related to the current paper. Koop and Potter (2004) use the factor structure framework of Stock and Watson (2002a) and propose the use of BMA to search over models which allow for non-sequential factors. Therefore resolving the issue of potentially selecting irrelevant factors when sequentially selecting the ones with highest eigenvalues. They apply BMA and Bayesian selection to the problem of forecasting GDP and inflation using quarterly data on 162 time series from 1959Q1 through 2001Q1. In their framework, the model allows for lags of the dependent variable and for factors extracted using principal component analysis. Since the factors are orthogonal, model search is performed over all indexes using the computationally algorithm introduced in Clyde (1999). Jacobson and Karlsson (2004) use BMA to find best predictors for the Swedish inflation. To traverse the model space, the number of predictors is limited to 20 (from an initial count of 80) and use a reverse jump Markov Chain Monte Carlo to search the model space.

Wright (2009) consider using BMA for pseudo out-of-sample prediction of US inflation. Using a pool of 107 predictors and quarterly data from 1960Q1 to 2006Q1, the study finds that BMA outperforms equal weighted model averaging. To overcome the dimensionality problem, Wright (2009) restricts the model space to only models with one predictor (i addition to lagged inflation). Hence, one contribution of the present paper is to adapt fast and efficient algorithms used in the Bayesian variable selection literature to search over large dimensional space. Instead of restricting the number of predictors in the forecasting equation as in Wright (2009), this paper uses the Markov Chain Monte Carlo algorithm developed by Kohn et al. (2001) to perform an "efficient" sweep of the model space even with large number of predictors. The algorithm reduces the computational burden by decreasing the algorithm visits to useless subsets of predictors and thus identifies "useful" models.

The literature on dimension reduction using principal components and Bayesian model averaging apparently moved in two different directions. However, recent findings by De Mol et al. (2008) and the (Ouysse and Kohn, 2009) suggest there are theoretical and practical reasons to connect the two literatures. De Mol et al. (2008) compare the properties of forecasts based on principal component regression, Ridge and lasso regressions, and find that these methods produce forecasts which are highly correlated with similar out-of-sample performance. They also consider double (N, T) asymptotics for the case of shrinkage regression with Gaussian prior. They find that consistency of the Bayesian (Ridge) regression forecast requires that the amount of shrinkage grows asymptotically at a rate equal to the number of predictors N. In the context of Bayesian variable selection, Ouysse and Kohn (2009) find that under empirical Bayes prior, more evidence is extracted from the data with a larger number of cross-sections and not necessarily from longer time series.

The main aim of the study is to compare the 'real time' out-of-sample forecast performance of two competing approaches for forecasting with high-dimensional panels: Bayesian model averaging and information aggregation using PC regression. Contrary to existing literature, the paper applies a fully Bayesian analysis and implements Bayesian variable selection over the full set of 131 predictors. The dataset employed is the same as the one used in De Mol et al. (2008) and Stock and Watson (2005) and comprises of monthly observations from 1959:01 to 2003:12 and 131 time series.

One contribution of this study is to compare the differences in the relative predictive performance between the predictors and their principal components. To this end we apply the same Bayesian analysis on all the principal components extracted from the full set of regressors. This allows non-sequential selection of the factors as in Koop and Potter (2004).

The single variable analysis that has been the main focus of the literature on forecasting imposes independence across output and inflation. This means the loss in output prediction errors is assumed to be independent of the loss in inflation prediction errors. Singly forecasting output-inflation may create situation in which the losses are compounded jointly (Komunjer and Owyang (2011)). This paper uses a Normal inverse-Wishart conjugate prior to estimate the forecasting equations of output and inflation in a system that allows the forecast errors to be correlated.

We conduct a different out-of-sample investigation in which the predictors are chosen jointly for both output and inflation using Bayesian variable selection in each out-of-sample recursion using information available at the time of the forecast in a ten years rolling window. In this framework, the posterior densities of the model and the parameters are time variant. This implies that the combining weights and the composition of the combined models are time varying. The "reduced" form time variation of the forecasting model is nonparametric which offers flexibility in capturing structural changes and instabilities of unknown forms.

The results show that in terms of mean-squared forecast error, principal components based forecasts have a slight marginal advantage over BMA. This edge of PCR in its global forecast performance hides important changes in the local forecasting power of the two approaches. The time varying profile of the BMA combining weights and thereby the profile of the posterior modal model further support the observation of existence of unstable environment. We use the fluctuation statistic of Giacomini and Rossi (2010) to assess the local out-of-sample relative performance of the competing forecasting models over the entire time path. There are instabilities in the forecasting performance of BMA and PC relative to the naive random walk and relative to each other. The profile of the relative local forecasting performance reveals surprising results. PC regression based forecasts performed generally worse than the random walk in the post 1985 period with high volatility for the late 70's and early 80's, where PC outperformed the naive benchmark. BMA on the other hand, performed consistently better than PC and the naive benchmark over the period prior to 1985. Thereafter, the performance of both BMA and PC was relatively modest compared to the naive benchmark. However, these profile differences are not statistically significant for industrial production. The significant differences in the profile performance is however significant for the consumer price index. On the other hand, PCR performed consistently worse than the random walk and BMA over the entire time path. BMA beats the naive benchmark for the period prior to 1985 and is better than PC. The differences in the relative performance between PC and both BMA and RW are statistically significant for the post 1985 period but only up to 2001.

Furthermore, an analysis of the Theil index (Theil (1967), Chauvet and Potter (2012)) indicates that the loss of performance of PCR is due mainly to its exuberant biases in matching the mean of the two series especially the inflation series. BMA forecasts series matches the first and second moments of the GDP and inflation series very well with practically zero biases and very low volatility.

BMA forecasting performance is robust to the choice of the hyperparameter of the g-prior. The prior affect the composition of the modal model and the combining weights. However, the informational value of these alternative BMA combinations is comparable, thereby suggesting a great deal of substitutability between models. This is perhaps due to the high level of correlation between the predictors variables. The profile performance of these methods does however change over time. Overall there is predictability in the late 70's and early 80's when the Fed exercised passive monetary policy and the US economy was subject to expectations-driven inflation. Predictability diminishes to some extent during the period of "Great moderation" starting in the late 80's when the Fed adapted an active monetary policy of inflation targeting.

2 Approaches to dimension reduction

2.1 Preliminaries

Using the notation and framework in De Mol et al. (2008), consider an $(n \times 1)$ vector of covariance stationary processes $Z_t = (z_{1t}, \cdots, z_{nt})'$ with 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, the aim is to estimate the linear projection, $\mathbf{y}_{t+h|t} = proj\{\mathbf{y}_{t+h}|\Im_t\}$, where $\Im_t = span\{Z_{t-s}, s = 0, 1, 2, \cdots\}$ is a potentially large information set, and $\mathbf{y}_{t+h} = (y_{1,t+h,\cdots,\mathbf{y}_{m,t+h}})$ is an $m-$vector of filtered versions of z_{it}, $y_{j,t+h} = f_{j,h}(L)z_{i,t+h}$ for specific $i = 1 \cdot \cdot, n$ and $1 \leq m \leq n$.

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

$$y_{j,t+h} = Z_t'\beta_{j,0} + \cdots + Z_{t-p}'\beta_{j,p} + u_{t+h} = X_t'\beta_j + u_{j,t+h},$$

where $\beta_j = (\beta_{j,0}, \cdots, \beta_{j,p})'$ and $X_t = (Z_t', \cdots, Z_{t-p}')$ for each target series j, $j = 1, \cdots, m$. Given a sample of size T, let $\mathbf{X} = (X_{p+1}, \cdots, X_{T-h})'$ be the $(T-h-p) \times n(p+1)$ matrix of observations for the predictors and $y_j = (y_{j,p+h+1}, \cdots, y_{j,T})'$ is the $(T-h-p) \times 1$ matrix of observations for the dependent variable. The traditional forecast is given by $\widehat{y}_{j,T+h|T}^{LS} = \mathbf{X}'\widehat{\beta}^{LS}$, where $\widehat{\beta}_j^{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y_j$, $j = 1, \cdots, m$.

When the size of the information set is large, this projection involves estimation of a large number of parameters, implying loss of degrees of freedom and poor forecasts. In addition, if $n \times (p+1) > T$, ordinary least squares is not feasible. Stock and Watson (2006) shows in the case of orthogonal regressors that the variance of the forecast error is proportional to N/T where $N = n \times (p+1)$ is the number of predictors. Therefore if, N is large relative to T, then the contribution of OLS estimation error to the forecast does not vanish, no matter how large the sample size.

2.2 The diffusion index framework

We consider forecasting situation in which both N and T are large, hence the double (N, T) asymptotics wit no requirements on the relative rates of convergence of N and T. The number of predictor series can be very large, often larger than the number of observations as it is the case in macroeconomic forecasting. Many studies have simplified the high-dimensional problem $(N > T)$ by modelling the covariability of the series (the target variables to be forecast and the predictor series) in terms of few number of unobserved factors. This literature predominately uses principal components analysis to estimate these common factors which are then used in forecasting. To be specific, we assume the following 'diffusion index' forecasting framework of Stock and Watson (2002a) where (X_t, y_{t+h}) admit a factor model representation with r common latent factors F_t

$$\begin{aligned} X_t &= \Lambda F_t + \xi_t & (1) \\ y_{j,t+h} &= \delta_j F_t + v_{j,t+h}, \; j = 1, \cdots, m, & (2) \end{aligned}$$

where $F_t = (f_{1t}, \cdots, f_{rt})'$ are $r-$dimensional stationary processes, ξ_t is an $N \times 1$ vector idiosyncratic disturbances and v_{t+h} is the forecast error. We follow De Mol et al. (2008) and make the following assumptons about the factors, the $N \times r$ matrix Λ of factors loadings, the forecasting equation (2) and the error terms (ξ_t, v_{t+h}).

Assumption 1 ***(Factor structure equation (1))***

(i) $EF_tF_t' = I_r$.

(ii) *Λ is a non-random matrix with full rank r for each N: for some $r \times r$ positive definite matrix D_Λ, $\|\Lambda'\Lambda/N - D_\Lambda\| \to 0$ as $N \to \infty$;*

(iii) *ξ_t are orthogonal to the factors F_t with covariance matrix $\mathrm{E}\,\xi_t\xi_t' = \Psi$ of full rank for all N;*

(iv) *Weak cross-sectional dependance: there exist $M > 0$ such that for all N, $t = 1, \cdot\cdot\cdot, T$,*

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(\xi_{it}\xi_{jt})| \leq M.$$

Assumption 2 ***(Forecasting equation (2))***

(i) *$v_{j,t+h}$ are orthogonal to X_t for each N nd $j = 1, \cdots, m$: $T^{-1}\sum_t F_t v_{j,t+h} \to 0$;*

(ii) *$T^{-1}\sum_t \mathbf{v}_{t+h}\mathbf{v}_{t+h}' \to \Sigma_{\mathbf{v}}$, where $\mathbf{v}_{t+h} = (v_{1,t+h}, \cdots, v_{m,t+h})'$ and $\Sigma_{\mathbf{v}}$ is an $m \times m$ positive definite matrix.*

(iii) *$|\delta_j| \leq \infty$ for $j = 1, \cdots, m$. (See Stock and Watson (2002a).)*

The factors F_t are unobserved and the number of common factors r is also unknown. There are several methods for determining the number of factors r. Stock and Watson (1998) develop a consistent estimator for r based on the fit of the forecasting equation (2). Bai and Ng (2002) use information criteria to penalize the sum of squared residuals in model (1) to construct consistent estimator for r.

Principal components regression (PCR) computes the forecasts as a projection on the first few principal components (Forni et al. (2005), Giannone et al. (2004), Stock and Watson (2002a,b)). Let $\widehat{F}_t$ be the $T \times r$ matrix of the first r principal components of the predictors $\mathbf{X}$ and let $\mathfrak{I}_t^f = span\{\widehat{f}_{1t}, \cdot\cdot\cdot, \widehat{f}_{rt}\}$ with $r \ll N$ be a parsimonious representation of the information set $\mathfrak{I}_t$. Following De Mol et al. (2008), let S_x be the sample covariance matrix of the predictors X, $S_x = \mathbf{X}'\mathbf{X}/(T-h-p)$ and consider the spectral decomposition of S_x: $S_xV = VD$ where $D = diag(d_1, \cdots, d_N)$ is a diagonal matrix with d_i corresponding to the i^{th} highest eigenvalue of S_x, and $V = (\nu_1, \cdots, \nu_N)$ is the matrix whose columns corresponds to the normalized eigenvectors of S_x. The normalized principal components are defined as :

$$\widehat{f}_{it} = \frac{1}{\sqrt{d_i}} v_i' X_t, \text{ for } i = 1, \cdots, N^*$$

where $N^* \leq N$ is the number of non-zero eigenvalues.

The principal component forecast is defined as:

$$y^{PC}_{j,T+h|T} = proj\{y_{j,T+h} | \Im^f_T\}. \tag{3}$$

Once the factors are estimated via principal component analysis (PCA), the projection is computed by OLS treating the factors as observed:

$$y^{PC}_{T+h|T} = \widehat{\theta}' \widehat{F}_T, \tag{4}$$

$$\widehat{\theta}_j = (\widehat{F}_T \widehat{F}_T')^{-1} \widehat{F}_T' y_j, \quad \widehat{F}_T = (\widehat{f}_{1T}, \cdot \cdot \cdot, \widehat{f}_{rT})'. \tag{5}$$

The literature has addressed the asymptotic properties of the principal components regression for N and T going to infinity. Bai and Ng (2002) established consistency of the estimated number of factors and of the PC estimates of the factors and factor loadings. Bai (2003) derived the asymptotic distributions of the estimated factor structure and Stock and Watson (2002a,b) established conditions under which the PCR forecasts converge to the optimal forecast. The main underlying requirement in these results is that the sources of common dynamics remain limited as the number of cross sections increases to infinity. To be specific, Assumption 1(i)-(ii) imply that each of the factors have a nontrivial contribution to the variance of X_t and provides identification of the factors up to a change of sign. Assumption 1(iii) implies an approximate factor structure in the sense of Chamberlain and Rothschild (1983) and allows for limited weak cross-section dependence which dies out as N goes to infinity. This ensures that the first eigenvectors (corresponding to r largest eigenvalues) of S_x behave as the first r eigenvectors of $\Lambda' F F \Lambda/(T-h-p))$, the component of the total variance driven by the common factors.

2.3 Shrinkage regression

Ridge regression and the *lasso* are classical approaches to shrinkage regression that penalize large coefficients:

$$\widehat{\beta_j}^{(\kappa)} = \text{argmin}_{\beta_j} \left\{ (y_j - \mathbf{X}\beta_j)'(y_j - \mathbf{X}\beta_j) + \lambda \sum_{k=1}^{N} |\beta^{(\kappa)}_{j,k}| \right\} \tag{6}$$

for some penalization parameter $\lambda \geq 0$. Choosing $\kappa = 2$ yields *ridge regression* where $\widehat{\beta_j}^{ridge} = (\mathbf{X'X} + \lambda I_N)^{-1} \mathbf{X} y_j$. Choosing $\kappa = 1$ yields the lasso (Tibshirani (1996)) which has no closed form solution but the entire path of λ can be obtained using the LARS algorithm (Efron et al. (2004)). Both of the ridge and lasso estimators can be interpreted as the posterior mode under a particular prior that assumes independence of the parameters. For ridge regression the prior is $\beta_j | \sigma^2_\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon \lambda)$; for the *lasso* it is an independent identically distributed Laplace (double exponential) $p(\beta_{j,k}|\sigma^2_\epsilon) = \frac{\lambda}{2\sigma_\epsilon} e^{-\lambda|\beta_{j,k}|/\sigma_\epsilon}$.

Large values of the penalty parameter λ cause the coefficients of $\widehat{\beta}_j^{(\kappa)}$ to be shrunk towards zero. PCR and Ridge regression give non-zero weight to all predictors. The Laplace prior puts more mass near zero and in the tails inducing either large or zero estimates of the regression coefficients. Therefore the lasso favors sparse regression coefficients instead of many fairly small coefficients as might result in the ridge regression.

De Mol et al. (2008) provide conditions under which the ridge forecast is consistent and converges to the unfeasible population forecast. They find that the prior should shrink increasingly all regression coefficients to zero as the number of predictors rises. Moreover, the shrinkage parameter λ must grow asymptotically at a rate equal to the number of predictors N (See *Corollary 1* in De Mol et al. (2008)).

3 Bayesian Model averaging

Using the notation in Ouysse and Kohn (2009), consider the econometric model

$$\mathbf{y} = (I_m \otimes \mathbf{X})\, \beta + \epsilon, \tag{7}$$

where, $\mathbf{y} = (y_1', \cdots, y_m')'$, $\beta = (\theta_1', \cdots, \theta_m')$, ϵ is an $m \times T$ vector of error terms, and I_m is an $m \times m$ identity matrix. The specification (7) enables the estimation and inference for the m variables to be forecast simultaneously as in a system of seemingly unrelated regression. Therefore any correlation across the idiosyncratic components is taken into account in the posterior inference and therefore allows for gains of efficiency.

Bayesian variable selection defines a selector vector $\gamma = \{\gamma_j, j = 0, \cdots, N\}$, where N is the total number of possible predictors in $\mathbf{X}$, and γ_j is a Bernoulli random variable that takes value one if predictor j is allowed in the forecasting model, and zero otherwise. Therefore $\gamma = \{\gamma_j, j = 0, 1, ..., N\}$ is a selector vector over the columns of $\mathbf{X} = (X_0, X_1, ..., X_N)$, where $X_0 = \iota_T$. Let $q_\gamma = \gamma_0 + \cdots + \gamma_N$ be the number of predictors (columns of $\mathbf{X}$) in model γ. Adopting this notation, we can write (7) under model γ as

$$\underset{mT\times 1}{\mathbf{y}} = \underset{mT\times mq_\gamma}{(I_m \otimes \mathbf{X}_\gamma)}\ \underset{mq_\gamma \times 1}{\beta_\gamma} + \underset{mT\times 1}{\epsilon}, \tag{8}$$

where the subscript γ indicates that only columns and elements with the corresponding γ element being 1 are included. Since γ is a binary sequence, the number of models to be evaluated is 2^N, which corresponds to a very large sample space for the empirical example we are treating in this paper with $N = 131$ and $2^N = 2.77 \times 10^{39}$ possible models.

3.1 Bayesian variable selection

In Bayesian analysis, model selection, estimation of the parameters and inference about γ are done simultaneously allowing for uncertainty about all model unknowns to be integrated out in the posterior inference. We consider a standard hierarchical Bayes prior:

$$p(\beta, \gamma, \Sigma) \;\; = \;\; p(\beta|\Sigma, \gamma) p(\Sigma|\gamma) p(\gamma). \tag{9}$$

A commonly used prior for γ is

$$p(\gamma) = \prod_{j=1}^{N} \pi^{\gamma_j}(1-\pi)^{(1-\gamma_j)},$$

with π prespecified. The number of factors q_γ in (7) thus follows a binomial distribution. We follow Fernandez et al. (2001) and choose $\pi = 0.5$ implying that $p(\gamma) = 2^{-N}$: so the expected model size is $N/2$ and the standard deviation is $\sqrt{N/4}$. This prior allows each variable to be in or out of the model independently with the same probability 1/2. If a smaller (bigger) value of π is prespecified, then smaller (larger) models are preferred a priori. To allow for the intercept term, γ_0 is fixed in all models to 1. Using a Normal inverse-Wishart conjugate prior, we implement Bayesian variable selection by specifying a g-prior for $\beta|\Sigma$ as $N(0, c\Sigma \otimes H_\beta)$. The covariance matrix H_β determines the amount of structure. It can be chosen to either replicate the correlation structure of the likelihood by setting $H_\beta = (\mathbf{X}'\mathbf{X})^{-1}$, this is also the g-prior recommended by Zellner (1986); or to weaken the covariance in the likelihood by setting, $H_\beta = I_N$, which implies that the components of β are conditionally independent. The tuning parameter c can be model and data dependent as in the empirical Bayes prior (EB), hence the notation $\widehat{c}_\gamma$. The larger the value of c, the more diffuse (flatter) is the prior over the region of plausible values of β. The value of c should be large enough to reduce the prior influence. However, excessively large values can generate a form of the Bartlett-Lindley paradox by increasing the probability on the null model as $c \to \infty$. There is an asymptotic correspondence between fixed choices of c and the penalized sum-of-squares (classical) information criteria, see George and Foster (2000) and Chipman et al. (2001). In univariate analysis, the case of $c = T$ corresponds to the so called *unit information prior* which has the same amount of information about β as that contained in one observation. This prior leads to Bayes factors with asymptotic behavior similar to the Bayesian information criterion (BIC). The *risk information prior* (RIC) is obtained for $c = N^2$ (Donoho and Johnstone (1994)). A conjugate g-prior with fixed $c \cong 3.92$ corresponds asymptotically to Akaike's AIC. As $c \to \infty$, the penalty for dimension goes to infinity and the model size goes to zero. Finally, George and Foster (2000) defines the data dependent local empirical Bayes prior

$$\widehat{c}_\gamma^{EB} = \max\{F_\gamma - 1, 0\}, \text{ where } F_\gamma = \frac{R_\gamma^2/q_\gamma}{(1-R_\gamma^2)/(T-1-q_\gamma)},$$

and R_γ^2 is the R-squared of the regression of $\mathbf{y}$ on the covariates of the model γ. See Ouysse and Kohn (2009) for an adaptation to the multivariate case.

The prior on the covariance of ϵ is a inverse-Wishart $\Sigma^{-1} \sim \mathcal{W}_m(\omega, \Phi^{-1})$ where Φ is an $m \times m$ scale parameter, $\omega > m+1$ is a shape parameter. We choose $\omega = m+2$ which reflects a minimum amount of prior information and $\Phi = \widehat{\Sigma} + s^2 I_m$, where $\widehat{\Sigma}$ is the maximum likelihood estimator for Σ in the regression of Y on $\mathbf{X}$ and s^2 is the sample variance in the pooled regression of $\mathbf{y}$ on $(I_m \otimes \mathbf{X})$. The posterior density of γ conditional on the observed excess returns is

$$p(\gamma|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\gamma, \mathbf{X})p(\gamma)}{\sum_\gamma p(\mathbf{y}|\gamma, \mathbf{X})p(\gamma)} \propto p(\gamma)p(\mathbf{y}|\gamma, \mathbf{X}), \tag{10}$$

where $p(\gamma)$ is the prior on γ and $p(\mathbf{y}|\gamma, \mathbf{X})$ is the marginal likelihood of the observed data under model γ, with

$$p(\mathbf{y}|\gamma, \mathbf{X}) = \int_{\Sigma}\int_{\beta} p(\mathbf{y}|\beta, \Sigma, \gamma, \mathbf{X}) p(\beta, \Sigma, \gamma) d\beta d\Sigma. \tag{11}$$

Let $Y = (y_1, \cdots, y_m)$, $D_\gamma = \left(\mathbf{X}_\gamma'\mathbf{X}_\gamma + H_\beta^{-1}\right)$ and $S_\gamma = Y'\left(I_T - X_\gamma D_\gamma^{-1} X_\gamma'\right) Y$. Further let $\widehat{\beta}_\gamma$ be the maximum likelihood estimator of β in model γ defined as $\widehat{\beta}_\gamma = \left[I_m \otimes \left(\mathbf{X}_\gamma'\mathbf{X}_\gamma\right)^{-1}\mathbf{X}_\gamma'\right]\mathbf{y}$, the full conditionals for the model parameters are as follows:

$$p\left(\gamma|\mathbf{y}, \mathbf{X}\right) \quad \propto \quad |H_\beta|^{-\frac{m}{2}} \left|\mathbf{X}'\mathbf{X} + H_\beta^{-1}\right|^{-\frac{m}{2}} |\Phi|^{\frac{\omega}{2}} |\Phi + S_\gamma|^{-\frac{(T+\omega)}{2}} \tag{12}$$

$$\Sigma^{-1}|\mathbf{y}, \gamma \quad \sim \quad \mathcal{W}_m(\omega + T, (S_\gamma + \Phi)^{-1}) \tag{13}$$

$$\beta|\mathbf{y}, \Sigma, \gamma \sim \mathcal{N}\left(\widetilde{\beta}_\gamma, \Sigma \otimes D_\gamma^{-1}\right), \quad \text{where} \quad \widetilde{\beta}_\gamma = \left(I_m \otimes D_\gamma^{-1}\mathbf{X}_\gamma'\mathbf{X}_\gamma\right)\widehat{\beta}_\gamma \tag{14}$$

For the conditionally dependent prior $H_\beta = (\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}$, the mean of $\widetilde{\beta}_\gamma$ the posterior density in (14) becomes $\widetilde{\beta}_\gamma = \eta_\gamma \widehat{\beta}_\gamma$ with $\eta_\gamma = \frac{c_\gamma}{1+c_\gamma}$. Therefore the posterior mean of β shrinks the maximum likelihood estimator $\widehat{\beta}_\gamma$ of model γ towards zero. The term η_γ can be interpreted as the relative importance or weight that is given to the sample information relative to the prior information. It also measures the amount of shrinkage implied by the choice of the tuning parameters.

For the conditionally independent prior $H_\beta = I_N$ the mean of $\widetilde{\beta}_\gamma$ the posterior density in (14) becomes

$$\widetilde{\beta}_\gamma = \left\{I_m \otimes \left[\mathbf{X}_\gamma'\mathbf{X}_\gamma + \frac{1}{c_\gamma}I_N\right]^{-1}\mathbf{X}_\gamma'\mathbf{X}_\gamma\right\}\widehat{\beta}_\gamma = I_m \otimes \left[\mathbf{X}_\gamma'\mathbf{X}_\gamma + \frac{1}{c_\gamma}I_N\right]^{-1}\mathbf{X}_\gamma'\mathbf{y}. \tag{15}$$

Note that if the target variables in $\mathbf{y}$ are predicted equation by equation, then the prior on β_j can be written as $\mathcal{N}(0, \sigma_{\epsilon_j}^2 c I_N)$. In this case, $m = 1$ in the expression (15) and the posterior mean of $\beta_j|y_j, \mathbf{X}, \gamma$ corresponds to the *ridge* solution $\widehat{\beta}_j^{ridge}$ with ridge penalization parameter $\nu = 1/c_\gamma$ and $c_\gamma \equiv \frac{\sigma_{\beta_j}^2}{\sigma_{\epsilon_j}^2}$ in De Mol et al. (2008). When there is no shrinkage ($\nu \to 0$), the ridge solution is the least squares estimator of β. The latter case corresponds to $c_\gamma \to \infty$, that is a prior with large variance and very little information about β.

If we apply the conditions for convergence of the Bayesian forecast to its population counterpart in De Mol et al. (2008)(Corolarry 1) to the case of data-dependent prior, we would require that

$$\inf_{N,T\to\infty} \frac{\min eig\left[c\sigma_{\epsilon_j}^2(\mathbf{X}'\mathbf{X})^{-1}\right]}{\|c\sigma_{\epsilon_j}^2(\mathbf{X}'\mathbf{X})^{-1}\|} \quad > \quad 0 \tag{16}$$

and,

$$\|c\sigma_{\epsilon_j}^2(\mathbf{X}'\mathbf{X})^{-1}\| = O(NT^{\frac{1}{2}+\alpha}), \qquad \text{where } 0 < \alpha < 1/2 \tag{17}$$

to ensure that all regression coefficients are shrunk to zero at the same asymptotic rate. From these two results we can derive the following condition on the prior c to ensure consistency of the data-dependent Bayesian forecast

$$c = O\left(\min eig\left[\mathbf{X}'\mathbf{X}\right] NT^{\frac{1}{2}+\alpha}\right) \qquad 0 < \alpha < 1/2 \tag{18}$$

3.2 Bayesian model averaging

Bayesian model averaging provides a formal way of handling inference in the presence of multiple competing models. In BMA the posterior distributions of quantities of interest are obtained as mixtures of the model-specific distributions weighted by the posterior model probabilities, Clyde (1999). This approach enables construction of posterior probability intervals that take into account variability due to model uncertainty and gives more reliable prediction than using a single model (Madigan and Raftery (1994)).

Suppose that θ is a quantity of interest that has similar interpretation in each model. The BMA posterior distribution of θ is a weighted average of its model specific posterior distributions, where the weights are the posterior model probabilities

$$p(\theta|\mathbf{y}) \;=\; \sum_{\gamma} p(\theta|\mathbf{y},\gamma)p(\gamma|\mathbf{y}), \tag{19}$$

and the BMA point estimate of θ is

$$\widehat{\theta}_{BMA} = \mathrm{E}\,(\theta|\mathbf{y}) = \sum_{\gamma} \mathrm{E}\,(\theta|\mathbf{y},\gamma)p(\gamma|\mathbf{y}). \tag{20}$$

The BMA estimate of the posterior predictive density of $\mathbf{y}_{t+h}$, conditional on information $\otimes_T \equiv \{\mathbf{y},\mathbf{X}\}$ is

$$p(\mathbf{y}_{T+h}|\mathbf{y},\mathbf{X}) \;=\; \sum_{\gamma} p(\mathbf{y}_{T+h}|\mathbf{y},\mathbf{X},\gamma)p(\gamma|\mathbf{y},\mathbf{X}). \tag{21}$$

The BMA forecast for $\mathbf{y}t+h$, defined as the expected value of the density in (21), is

$$\widehat{\mathbf{y}}^{BMA}_{T+h|T} \;=\; \sum_{\gamma}(I_m \otimes \mathbf{X}_{\gamma})\widetilde{\beta}_{\gamma}p(\gamma|\mathbf{y},\mathbf{X}). \tag{22}$$

The BMA forecast (22) is the *optimal* Bayesian predictor of y_{T+h} in terms of expected loss over the posterior $p(\gamma|\mathbf{y},\mathbf{X})$ (Barbieri and Berger (2004)). Let R_{γ} be a diagonal matrix with diagonal element $R_{jj} = \gamma_j$ for $j = 1,\cdots,N$, then $\mathbf{X}_{\gamma} \equiv R_{\gamma}\mathbf{X}$ and we can rewrite (22):

$$\widehat{\mathbf{y}}^{BMA}_{T+h|T} \;=\; (I_m \otimes \mathbf{X})\sum_{\gamma} p(\gamma|\mathbf{y},\mathbf{X})R_{\gamma}\widetilde{\beta}_{\gamma}. \tag{23}$$

Implementation of (20) and therefore (22) is difficult because the sum over the 2^N possible models is impractical when N is large. One approach to get around this difficulty is to use MCMC and the simulated Markov chain from the posterior distribution $p(\gamma|\mathbf{y})$; $\gamma^{(j)}, j = 1,...,M$. Under suitable regularity conditions (Smith and Roberts (1993)), the posterior mean

$$\widehat{\theta}_{pm} = \frac{1}{M}\sum_{j=1}^{M} \mathrm{E}\,(\theta|\gamma^{(j)},\mathbf{y}), \tag{24}$$

is a consistent estimate of $E\,(\theta|\mathbf{y})$. We use the posterior mean estimate

$$\widehat{\beta}_{pm} \;=\; \frac{1}{M}\sum_{j=1}^{M}\beta^{(j)},$$

to approximate the BMA estimate of β. Similarly, the posterior mean estimate $\widehat{\Sigma}_{pm}$ is obtained as the sample mean of the MCMC draws $\Sigma^{(j)}, j = 1, .., M$. The BMA estimates of any function of γ are obtained by calculating the appropriate function at each draw and averaging. The quantity in (22) is approximated by the posterior mean forecast

$$\widehat{\mathbf{y}}^{pm}_{T+h|T} \quad = \quad \frac{1}{M} \sum_{j=1}^{M} (I_m \otimes \mathbf{X}_{\gamma^{(j)}}) \widetilde{\beta}_{\gamma^{(j)}}, \tag{25}$$

where $\widetilde{\beta}_{\gamma^{(j)}}$ is determined using (14). The matrix $\mathbf{X}_{\gamma^{(j)}}$ is formed by selecting the columns k of $\mathbf{X}$ corresponding to $\gamma^{(j)}_k = 1$, in other terms $\mathbf{X}_{\gamma^{(j)}} = R_{\gamma^{(j)}}\mathbf{X}$.

Barbieri and Berger (2004) show that under some regularity conditions, the optimal model for prediction under squared error loss is the model γ that minimizes

$$L(\gamma) = \left(R_\gamma \widetilde{\beta}_\gamma - \overline{\beta} \right)' Q \left(R_\gamma \widetilde{\beta}_\gamma - \overline{\beta} \right),$$

where $\overline{\beta} = \sum_\gamma p(\gamma|\mathbf{y}, \mathbf{X}) R_\gamma \widetilde{\beta}_\gamma$. These regularity conditions include $Q = a\mathbf{X}'\mathbf{X}$ for $a > 0$ and $\widetilde{\beta}_\gamma = b\widehat{\beta}_\gamma$. Both conditions satisfied in the conjugate *g-type* priors we consider in this section. Under these conditions, the model that minimizes $L(\gamma)$ is the *median* probability model consisting of those variables l whose posterior inclusion probability p_l is at least zero. The BMA estimate of the posterior (marginal) inclusion probability (probability that a predictor l is informative) is $\widehat{\pi}_l = \sum_{j=1}^{M} \gamma^{(j)}_l / M$, and the posterior average number of "informative" predictors in the model is denoted $\widehat{q}_{pm} = \sum_{j=1}^{M} \widehat{q}^{(j)}_\gamma / M$, where $\widehat{q}^{(j)}_\gamma = \sum_{l=1}^{N} \gamma^{(j)}_l$.

Since we consider forecasting using rolling over estimates with a window of ten years, then at each time T BMA and Bayesian variable selection is carried out using the last 10 years of data. In our application T is rolled over from $T_0 = 1969 : 12$ to $T_1 = 2002 : 12$. At the end we have a time series of estimates of the parameters of the model. In particular, we are interested in the profile of the posterior median model, the distribution of the average size of the forecasting model and the distribution of the inclusion probabilities of the predictors. Therefore, we denote $\widehat{q}_{pm,T}$ and $\widehat{\pi}_{j,T}$ the BMA estimates of the average size of the forecasting model and the inclusion probability for X_j using data up to time T and we use the notation $\widehat{\mathrm{E}}\left(\widehat{q}_{pm,T}\right)$ and $\overline{\pi}_j$ to denote their sample average respectively over the forecasting period, where

$$\widehat{\mathrm{E}}\left(\widehat{q}_{pm}\right) \quad = \quad \sum_{T=T_0}^{T_1} \widehat{q}_{pm,T}/s \tag{26}$$

$$\overline{\pi}_j \quad = \quad \sum_{T=T_0}^{T_1} \widehat{\pi}^{pm}_{j,T}/s \tag{27}$$

and s is the number time periods in the evaluation period, $s = T_1 - T_0 + 1$.

4 Comparison of BMA and PC forecasts

The data series we use is the same as the one used in De Mol et al. (2008) and Stock and Watson (2005). The total number of predictors $N = 131$ in $\mathbf{X}$ includes

Table 1: Correlation of BMA out-of-sample forecasts of industrial production with Lasso, Ridge and PC.

Forecast period 1970 : 12 **to** 2002 : 12									
			LASSO with BMA						
			Number of non-zero coefficients						
	1	3	5	10	25	50	75		$\widehat{\mathrm{E}}\ (\widehat{q}_{pm})$
$c_\gamma = T$	0.4271	0.7422	0.8047	0.8623	0.8449	0.7782	0.6132		7.25
$c_\gamma = N^2$	0.5009	0.8194	0.8539	0.8475	0.7801	0.692	0.5131		2.55
$c_\gamma = 4$	0.4998	0.7496	0.8006	0.8672	0.9072	0.9132	0.8038		32
			RIDGE with BMA						
			In sample residual variance, κ						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν	6	25	64	141	292	582	1141	2339	6025
$c_\gamma = T$	0.6519	0.7767	0.8146	0.8231	0.8153	0.7914	0.7436	0.6445	0.4107
$c_\gamma = N^2$	0.5716	0.7323	0.7992	0.8285	0.836	0.8239	0.7858	0.695	0.4631
$c_\gamma = 4$	0.8448	0.9056	0.8982	0.8764	0.8502	0.8188	0.7749	0.6947	0.5027
			PC with BMA						
			Number of principal components, r						
	1	3	5	10	25	50	75		
$c_\gamma = T$	0.2139	0.7218	0.776	0.7954	0.7937	0.7278	0.6076		
$c_\gamma = N^2$	0.2577	0.7699	0.8178	0.8363	0.8011	0.664	0.5014		
$c_\gamma = 4$	0.1644	0.6924	0.7188	0.7592	0.7955	0.8205	0.7158		
Forecast period 1970 : 12 **to** 1984 : 12									
			LASSO with BMA						
			Number of non-zero coefficients						
	1	3	5	10	25	50	75		
$c_\gamma = T$	0.3312	0.7598	0.8397	0.9045	0.8898	0.8248	0.6569		
$c_\gamma = N^2$	0.4793	0.8666	0.9099	0.8928	0.8263	0.7397	0.5568		
$c_\gamma = 4$	0.393	0.7562	0.8175	0.9008	0.9402	0.9338	0.825		
			RIDGE with BMA						
			In sample residual variance, κ						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν	6	25	64	141	292	582	1141	2339	6025
$c_\gamma = T$	0.6946	0.8297	0.865	0.8696	0.8582	0.8301	0.7756	0.6595	0.3613
$c_\gamma = N^2$	0.6251	0.7955	0.8567	0.8793	0.882	0.8676	0.8296	0.737	0.4697
$c_\gamma = 4$	0.8681	0.9339	0.9252	0.9027	0.8753	0.8406	0.789	0.689	0.431
			PC with BMA						
			Number of principal components, r						
	1	3	5	10	25	50	75		
$c_\gamma = T$	0.1253	0.7591	0.8346	0.8495	0.8393	0.7611	0.6278		
$c_\gamma = N^2$	0.1909	0.8005	0.8608	0.8785	0.8462	0.703	0.5284		
$c_\gamma = 4$	0.0687	0.7323	0.7701	0.8084	0.8392	0.8332	0.7437		
Forecast period 1985 : 01 **to** 2002 : 12									
			LASSO with BMA						
			Number of non-zero coefficients						
	1	3	5	10	25	50	75		
$c_\gamma = T$	0.6877	0.7771	0.7797	0.7928	0.762	0.7053	0.555		
$c_\gamma = N^2$	0.6542	0.732	0.7086	0.702	0.6371	0.585	0.4381		
$c_\gamma = 4$	0.706	0.7677	0.7943	0.8194	0.855	0.8813	0.7669		
			RIDGE with BMA						
			In sample residual variance, κ						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν	6	25	64	141	292	582	1141	2339	6025
$c_\gamma = T$	0.5741	0.6654	0.71	0.7381	0.7537	0.7547	0.7383	0.6984	0.618
$c_\gamma = N^2$	0.4428	0.5414	0.6001	0.6439	0.675	0.6891	0.6818	0.6458	0.5623
$c_\gamma = 4$	0.7992	0.8559	0.867	0.8627	0.8482	0.8239	0.7902	0.7444	0.6733
			PC with BMA						
			Number of principal components, r						
	1	3	5	10	25	50	75		
$c_\gamma = T$	0.554	0.7276	0.6659	0.6722	0.6949	0.6639	0.5547		
$c_\gamma = N^2$	0.5779	0.7013	0.631	0.6459	0.6025	0.5465	0.4307		
$c_\gamma = 4$	0.4858	0.6731	0.6915	0.7236	0.7544	0.8082	0.6768		

Figure 1: Industrial production 12− step ahead out-of-sample forecasts.

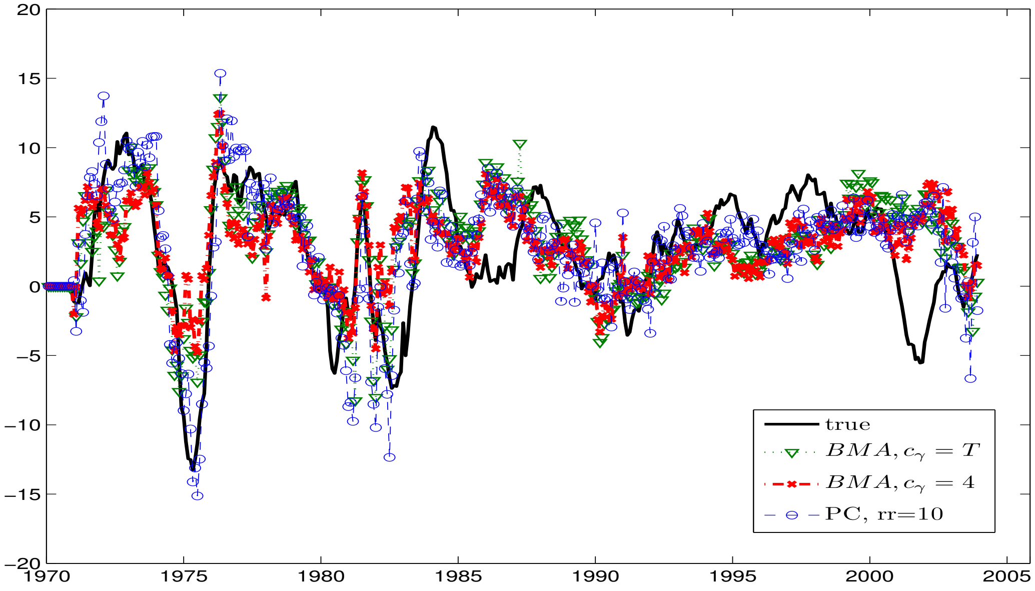
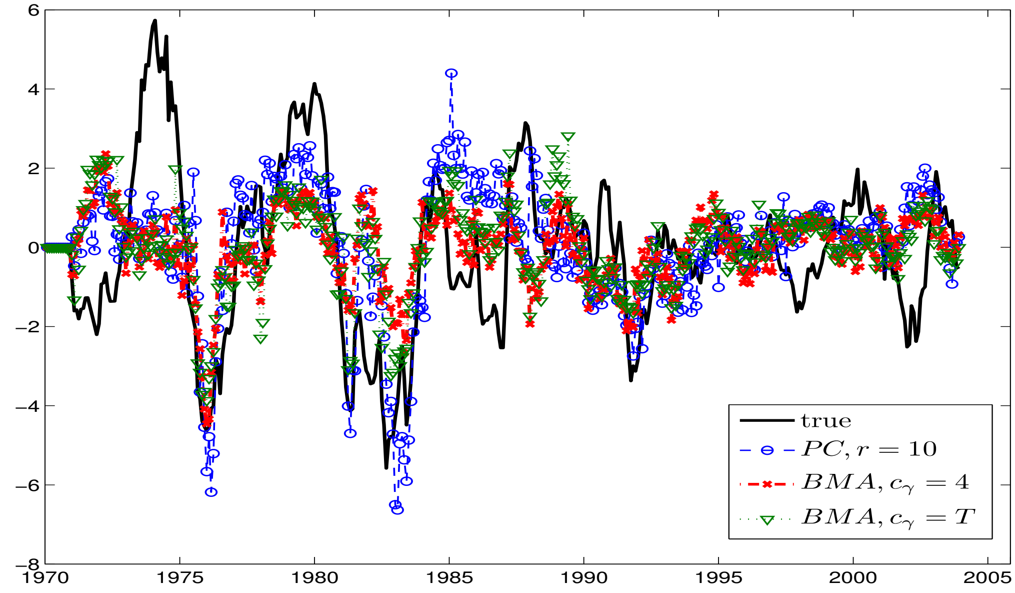


Figure 2: Consumer Price Index 12− step ahead out-of-sample forecasts.



real variables such as sectoral industrial production, employment and hours worked; nominal variables such as consumer and price indexes, wages, money aggregates; in addition to stock prices and exchange rates. The data series are transformed to achieve stationarity: monthly growth rates for real variables(industrial production, sales $\cdots$) and first differences for variables already expressed in rates (unemployment rate, capacity utilization, $\cdots$).

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

$$\begin{aligned} z^h_{IP,t+h} &= (ip_{t+h} - ip_t) = z_{IP,t+h} + \cdots + z_{IP,t+1} \\ z^h_{CPI,t+h} &= (\pi_{t+h} - \pi_t) = z_{CPI,t+h} + \cdots + z_{CPI,t+1} \end{aligned}$$

$IP_T = 100 \times \log IP_t$ is the rescaled logarithm of IP, $cpi_t = 100 \times \log \frac{CPI_t}{CPI_{t-12}}$ IP enters the 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:

$$\begin{aligned} \widehat{ip}_{T+h|T} &= \widehat{z}^h_{IP,T+h|T} + ip_T \\ \widehat{cpi}_{T+h|T} &= \widehat{z}^h_{CPI,T+h|T} + cpi_T \end{aligned}$$

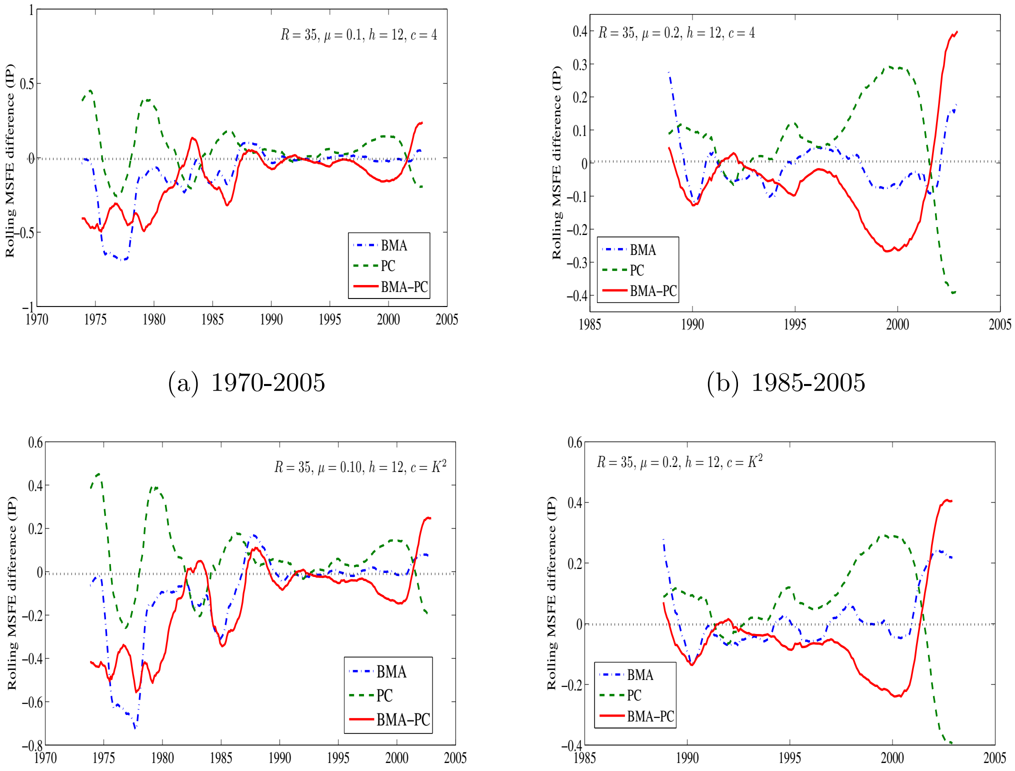
In this section we compare the performance of BMA forecasts to those based on principal components and shrinkage (ridge and lasso) regression. Figure 6 and Table 1 show the sample correlation among BMA forecasts and Ridge forecasts $\widehat{\rho}_{Ridge}$, among BMA forecasts and lasso forecast $\widehat{\rho}_{lasso}$, and among BMA forecasts and principal components forecasts $\widehat{\rho}_{PC}$. The PCR forecasts depend on the number of factors allowed in the factor structure 1. Similarly, the Ridge and lasso regression forecasts depend on the choice of the regularization parameter λ in 6. We follow De Mol et al. (2008) and report sample correlation for $r = 1, 3, 5, 10, 25, 50, 75$. For the Ridge regression, the priors are chosen for which the in-sample fit explains a given fraction $1 - \kappa$ of the variance of the variable to be forecast. For the Lasso, the prior on β is selected to deliver a given number ($= r$) of non-zero coefficients.

4.1 Measures of out-of-sample performance

The dataset employed is the same as the one used in De Mol et al. (2008) and Stock and Watson (2005) and comprises of monthly observations from 1959:01 to 2003:12 and 131 time series. The sample is divided into an in-sample portion of size $T = 120$ (1959:01 to 1969:12) and an out-of-sample evaluation portion with first date December 1970 and last date December 2003. Therefore, there are a total of $M = 397$ out-of-sample evaluation points which we split into pre- and post-1985 periods with cat-off date December 1984. The models and parameters are re-estimated and the 12-step-ahead forecasts are computed for each month $t = T + 12, \cdots T + 12 + M - 1$ using a rolling window scheme that uses the most recent 10 years of monthly data, that is data indexed $t - 12 - T + 1, \cdots, t - 12$.

Let $\widehat{e}_{t|t-12}$ be the forecast error computed at time t, $\widehat{e}_{t|t-12} = y_t - \widehat{y}_{t|t-12}$, where $\widehat{y}_{t|t-12}$ is the computed point forecast. One measure of overall average performance is the square root of the out-of-sample mean square forecast error (RMSFE) calculated

Figure 3: Fluctuation test statistic for industrial production, obtained as the standardized difference between the MSFE of the PC regression model and the MSFE of the random walk (PC), between the MSFE of the BMA and the MSFE of the random walk (BMA), and between the MSFE of BMA and the PC regression (BMA-PC)



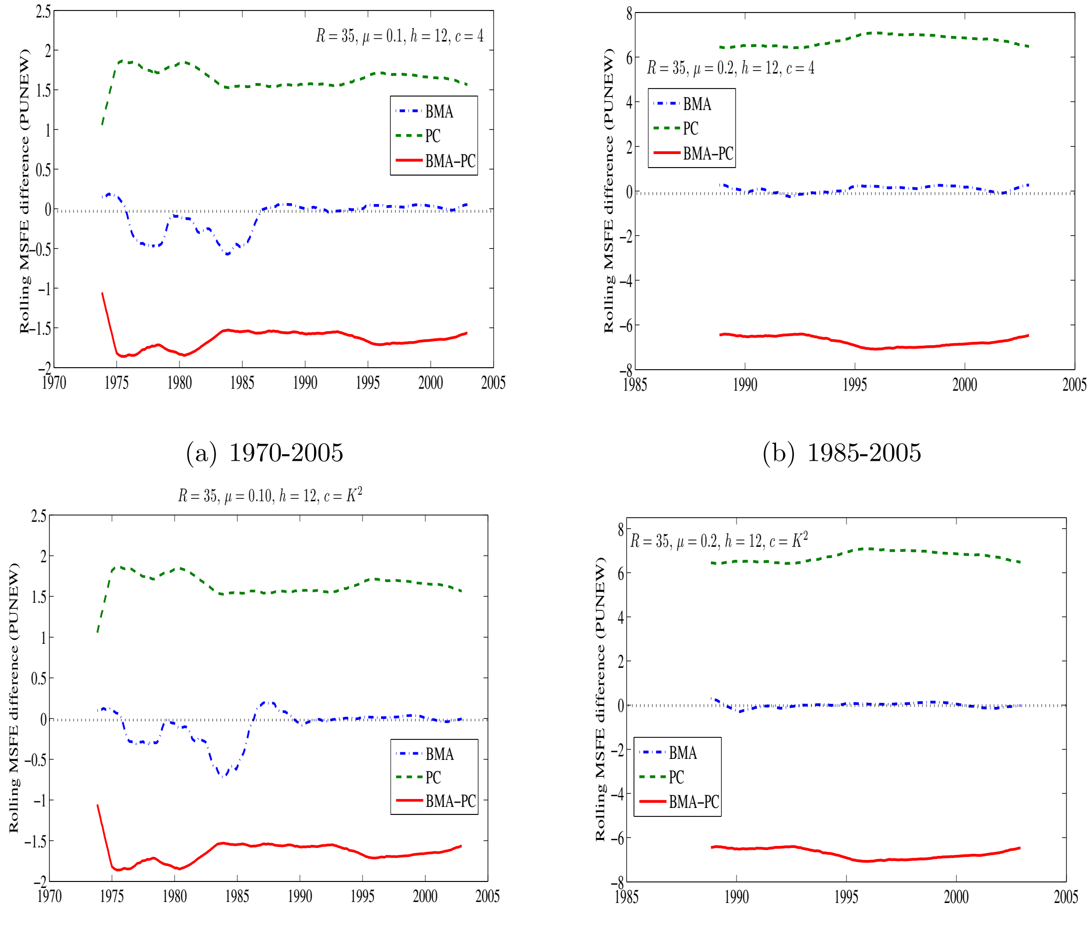
(a) 1970-2005

(b) 1985-2005

(c) 1970-2005

(d) 1985-2005

Figure 4: Fluctuation test statistic for consumer price index (CPI), obtained as the standardized difference between the MSFE of the PC regression model and the MSFE of the random walk (PC), between the MSFE of the BMA and the MSFE of the random walk (BMA), and between the MSFE of BMA and the PC regression (BMA-PC)



(a) 1970-2005

(b) 1985-2005

(c) 1970-2005

(d) 1985-2005

Table 2: Root mean-squared forecast errors relative to the naive random walk model.

Industrial Production											
	Bayesian model averaging								Principal Component		
Forecast	$c_\gamma = T$		$c_\gamma = 4$		$\widehat{c}_\gamma^{EB}$		$c_\gamma = N^2$		r		
Period	X	F	X	F	X	F	X	F	5	10	25
$1971-2002$	0.65	0.87	0.63	0.86	0.63	0.86	0.63	0.87	0.56	0.54	0.65
	(0.64)	(0.66)	(0.75)	(0.51)	(0.77)	(0.49)	(0.76)	(0.50)	(0.79)	(0.97)	(1.28)
$1971-1984$	0.40	0.56	0.40	0.69	0.39	0.69	0.39	0.70	0.35	0.34	0.46
	(0.59)	(0.54)	(0.79)	(0.47)	(0.79)	(0.44)	(0.79)	(0.45)	(0.93)	(1.11)	(1.43)
$1985-2002$	1.39	1.81	1.28	1.350	1.36	1.36	1.31	1.36	1.16	1.13	1.21
	(0.78)	(1.01)	(0.64)	(0.62)	(0.72)	(0.63)	(0.64)	(0.62)	(0.33)	(0.51)	(0.79)
Consumer Price Index											
	Bayesian model averaging								Principal Component		
Forecast	$c_\gamma = T$		$c_\gamma = 4$		$\widehat{c}_\gamma^{EB}$		$c_\gamma = N^2$		r		
Period	X	F	X	F	X	F	X	F	5	10	25
$1971-2002$	0.70	0.78	0.72	0.77	0.74	0.80	0.72	0.777	0.57	0.69	0.83
	(0.39)	(0.50)	(0.33)	(0.52)	(0.33)	(0.53)	(0.37)	(0.52)	(0.61)	(0.63)	(0.69)
$1971-1984$	0.56	0.67	0.57	0.68	0.62	0.71	0.59	0.68	0.39	0.48	0.56
	(0.34)	(0.49)	(0.31)	(0.49)	(0.30)	(0.50)	(0.33)	(0.50)	(0.57)	(0.57)	(0.60)
$1985-2002$	1.34	1.29	1.41	1.23	1.28	1.23	1.34	1.22	1.43	1.71	2.11
	(0.54)	(0.53)	(0.44)	(0.61)	(0.43)	(0.61)	(0.48)	(0.59)	(0.73)	(0.83)	(0.95)

The results under X correspond to those obtained from applying BMA to the model in (7) with the predictors X. The results under F correspond to BMA applied to the factors $\widehat{F}$ estimated by extracting the principal components of X as in Equation (34).

as

$$RMSFE = \sqrt{\sum_{j=T+12}^{T+12+M-1} \widehat{e}^2_{j|j-12}/M}$$

The mean square forecast error provides an estimate of the average performance over the whole out-of-sample evaluation period. This measure has been shown to perform poorly in the presence of instabilities. Giacomini and Rossi (2010) give the example of forecasting the dollar/British pound exchange rate and find that despite that the RMSFE for the random walk model is smaller than that for the uncovered interest rate parity model, the relative performance of the two models changes considerably over the sample. The authors highlight the fact that global relative forecasting performance may hide important information about the relative performance over time. The authors propose alternative measures for the evolution of the relative forecasting performance of competing models, one of which is the Fluctuation test. The Fluctuation test statistic, $F_{t,l}$ defined as:

$$F_{t,l} \;=\; \widehat{\sigma}^{-1} l^{-1} \sum_{j=t-l/2}^{t+l/2-1} \Delta \widehat{L}_j^{s_1,s_2}, \tag{28}$$

where $\Delta \widehat{L}_j^{s_1,s_2}$ is the difference in forecast error between models s_1 and s_2,

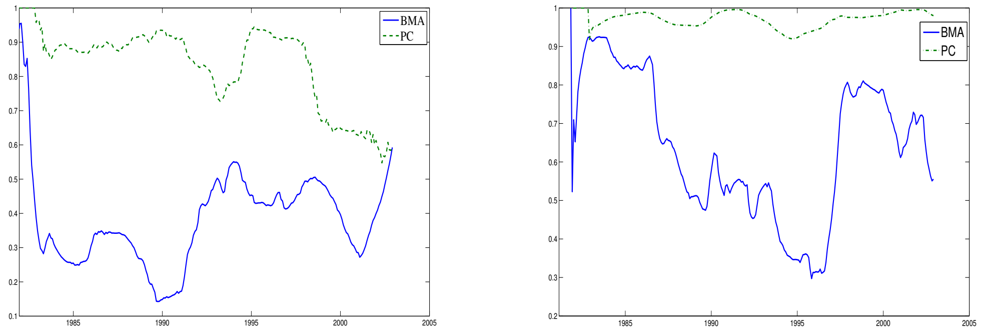
$$\Delta \widehat{L}_j^{s_1,s_2} = \widehat{e}_j^{2(s_1)} - \widehat{e}_j^{2(s_2)}.$$

The fluctuation test statistic is a difference between the MSFE of models s_1 and s_2 calculated over rolling window of size l. The null hypothesis of the test is that there is no difference in the (local) relative out-of-sample performance of models s_1 and s_2.

Consider the case of industrial production, Figure 3 shows the BMA and PCR local relative performance computed over rolling windows of $R = 35$. The figure shows three fluctuation statistics, denoted BMA, PC and BMA-PC. BMA (resp. PC) is the statistic $F_{t,l}$ with model s_1 being the BMA (resp. PC) and s_2 being the random walk. The third statistic $BMA - PC$ is computed as out-of-sample MSFE differences between BMA and PCR, that is statistic $F_{t,l}$ with s_1 is BMA and s_2 is PC. Negative values of the statistic indicate better local relative performance of model s_1 over s_2. Overall the statistic fluctuates between -0.5 and 0.5, which is statistically insignificant using the reported critical values in Giacomini and Rossi (2010). Quantitatively however, BMA produced better out-of-sample forecasts than the random walk model. What is surprising and novel is the performance of PC and BMA. The latter outperforms PC regression over the evaluation sample especially for the the pre 1985 era. PC even fails to beat the random walk over some periods of the evaluation sample. The other key observation is that both PC and BMA provide little advantage over the random walk for the 90's with a drastic fall in their performance post 2001.

For the consumer price index, Figure 4 shows that while the local performance of BMA relative to random walk is marginal, BMA produced better forecasts than PC. The latter consistently produced consistently worse forecasts than the random walk and BMA. Another measure of discrepancy between forecasts and the actual values is

Figure 5: Time-Varying Theil Index over the out-of-sample forecast evaluation period $1970 : 12 - 2002 : 12$ computed over a rolling window of $R = 35$.



(a) Industrial Production, 1970-2005

(b) CPI, 1970-2005

the Theil inequality index (Theil (1967)). Theil index was originally proposed for the measurement of income inequality and is an application of the concept of conditional entropy to the measurement of distributional change. Theil index has a decomposable structure allowing the additive disaggregation of the index in three terms, respectively related to bias, variance and covariance measures. The U-index proposed by Theil (1967) and applied to the context of forecast evaluation over the entire evaluation sample is given by the expression,

$$U = \frac{\sqrt{\frac{1}{M}\sum_{j=T+12}^{T+12+M-1}(y_j - \hat{y}_{j|j-12})^2}}{\sqrt{\frac{1}{M}\sum_{j=T+12}^{T+12+M-1} y_j^2} + \sqrt{\frac{1}{M}\sum_{j=T+12}^{T+12+M-1} \hat{y}_{j|j-12}^2}}. \qquad (29)$$

The U-index is scale invariant and fluctuates between zero and one, where zero indicates perfect forecast. The index also has an additive decomposition into three components:

Table 3: Theil Inequality Index and its decomposition into percentage of bias (%Bias), variance (% Var) and covariance (% Cov) for BMA and PC forecasts. The forecast evaluation period is 1970 : 12 − 2002 : 12. Values in parenthesis are computed over the recession periods as defined in the NBER dating.

Industrial Production					Consumer Price Index			
	PC, $r = 10$				PC, $r = 10$			
	Theil	% Bias	% Var	%Cov	Theil	% Bias	% Var	%Cov
	0.784	0.253	0.275	0.473	0.973	0.960	0.028	0.012
	(0.849)	(0.119)	(0.273)	(0.624)	(0.969)	(0.996)	(0.000)	(0.003)
	BMA				BMA			
c_γ	Theil	% Bias	% Var	%Cov	Theil	% Bias	% Var	%Cov
4	0.641	0.003	0.305	0.694	0.432	0.010	0.187	0.804
	(0.696)	(0.368)	(0.042)	(0.600)	(0.807)	(0.040)	(0.569)	(0.406)
K^2	0.376	0.034	0.022	0.945	0.581	0.009	0.179	0.813
	(0.593)	(0.134)	(0.005)	(0.875)	(0.633)	(0.104)	(0.382)	(0.528)
T	0.376	0.034	0.022	0.945	0.581	0.009	0.179	0.813
	(0.599)	(0.218)	(0.000)	(0.794)	(0.707)	(0.069)	(0.588)	(0.358)

bias, variance and covariance. The proportion of these three components are given by the expressions (Chauvet and Potter (2012)),

$$\%Bias = \frac{\left[\frac{1}{M}\sum_{j=T+12}^{T+12+M-1}\hat{e}_{j|j-12}\right]^2}{\sum_{j=T+12}^{T+12+M-1}\hat{e}^2_{j|j-12}/M} \tag{30}$$

$$\%Var = \frac{\left(\sqrt{V(\hat{y}_{j|j-12})}-\sqrt{V(y_j)}\right)^2}{\sum_{j=T+12}^{T+12+M-1}\hat{e}^2_{j|j-12}/M} \tag{31}$$

$$\%Cov = 1-\%Bias-\%Var \tag{32}$$

These components compare the moments of the forecasts to those of the actual data. In particular, the Bias and variance proportions are measures of departure of the mean and the variance of the forecasts from those of the actual series. Therefore, smaller bias and variance proportions are desirable meaning that the largest component in the Theil index comes from the covariance proportion.

Figure 5 shows the fluctuation over time of the Theil index computed over the entire out-of-sample evaluation period using a rolling window $R = 35$. The inequality index for the PC regression forecasts is significantly high with values exceeding 60% for industrial production and no less than 95% for consumer price index. This indicates that there is a significant gap between the PC regression forecasts series and the actual series. BMA forecasts perform much better with values lower than 20% for GDP and less than 30% for inflation in early and mid 1990's.

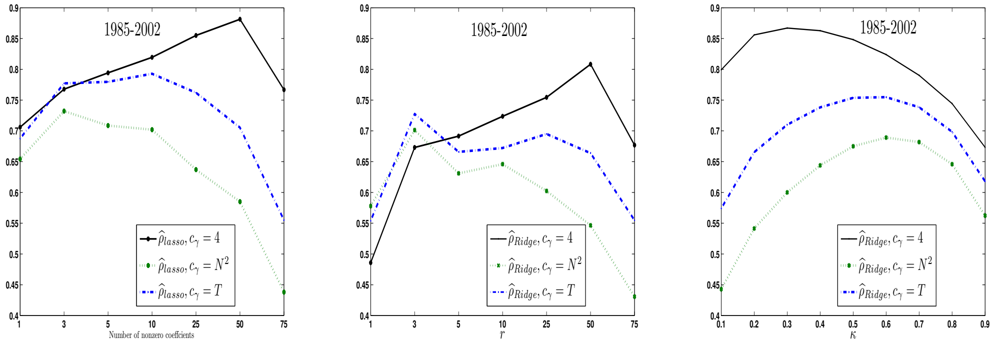
Table 3 reports the results of the decomposition of the Theil index in its three components over the entire out-of-sample evaluation period and over the recession periods (in parenthesis). The results highlight a very important point. Although the overall measure of discrepancy of PC indicates worse performance than BMA, there is a difference in how well the two models capture the moments of the actual series.

Remarkably, the prior on t he parameter c does affect the ability of the BMA forecasts to track the volatility of the GDP series. BMA with unit and risk information prior displays remarkable ability in forecasting the volatility of GDP series with variance proportion as low as 2.2% while the Akaike prior $c = 4$ results in higher variance component of 30.5%. The biases of these forecasts are substantially low with 3.4% for the former and 0.3% for the latter. This result is consistent with the finding (Table 2) that BMA forecasts under $c = 4$ have low variance compared to those generated from c equal to K^2 and T. The PC regression forecasts series does very pour both in matching the mean and the variance of the actual series with bias proportion of 25.3% and variance proportion of 27.5%. In terms of the consumer price index, the PC forecasts display good forecasting accuracy of the volatility of the actual series with variance proportion of (2.8%) compared to the BMA forecast which a variance proportion in the range of 18%. The PC low volatility proportion comes however at a high bias of 96% while BMA forecasts have very low to almost no bias for all prior settings.

In general the variance proportion is smaller over the recession period indicating that the forecast series become more volatile and thus is able to capture more of the variance of the series. The bias proportion generally increases during recession.

The analysis of the relative performance of the competing PC and BMA models shows that there is some gains from using BMA but mostly these are marginal and not statistically significant. This suggest that there may be common behaviour in these series. We turn now to the analysis of correlation between forecasts from the competing models. The following patterns can be seen in Figure 6. First, a *ranking* of the sample correlation with respect to the choice of the tuning parameter c_γ is apparent especially for the shrinkage based forecasts. The sample correlation is highest or at least reaches a maximum for $c_\gamma = 4$, followed by the case of $c_\gamma = T$. The sample correlation when $c_\gamma = N^2$ comes last. This means that the more informative the priors (therefore more shrinkage towards zero) the higher is the correlation between the forecasts generated by BMA and the three methods. Second, for $c_\gamma = 4, T$ the maximum correlation between the lasso forecasts and BMA is the highest compared to Ridge and PCR. Third, for lasso and PCR, the maximum correlation with BMA forecasts is reached at the same abscissa, that is for number of nonzero coefficients equal to the number of principal components allowed in the model. This number tends to be small $(= 3, 5)$ for $c_\gamma = t, N^2$ and large $(= 50)$ for $c_\gamma = 4$. For the Ridge forecasts in Figure 6(c), we see the opposite with maximum correlation reached at high values of κ $(= 0.6)$ for $c_\gamma = T, N^2$ and low values $(= 0.3)$ for $c_\gamma = 4$.

Figure 6: Sample correlation for BMA with Ridge ($\widehat{\rho}_{Ridge}$), PC ($\widehat{\rho}_{PCR}$) and Lasso ($\widehat{\rho}_{lasso}$) forecasts



(a) BMA and Lasso

(b) BMA and PCR

(c) BMA and Ridge

Table 1 further shows that these patterns generally hold for the full sample and the two subperiods. Under the priors $c_\gamma = T$ and $c_\gamma = N^2$, the sample correlation $\widehat{\rho}_{lasso}$ and $\widehat{\rho}_{PC}$ reach a maximum at the same values of r (10 and 5 respectively). Under the prior $c_\gamma = 4$, the highest correlation between BMA and lasso is reached for number of nonzero coefficient equal to 25 while the correlation of BMA and PC forecasts is at its maximum for $r = 50$. The BMA and ridge correlation $\widehat{\rho}_{ridge}$ is highest for $\kappa = 0.5$ and $\nu = 292$ when $c_\gamma = N^2$, $\kappa = 0.4$ and $\nu = 141$ for $c_\gamma = T$, and $\kappa = 0.2$ and $\nu = 25$ for $c_\gamma = 4$.

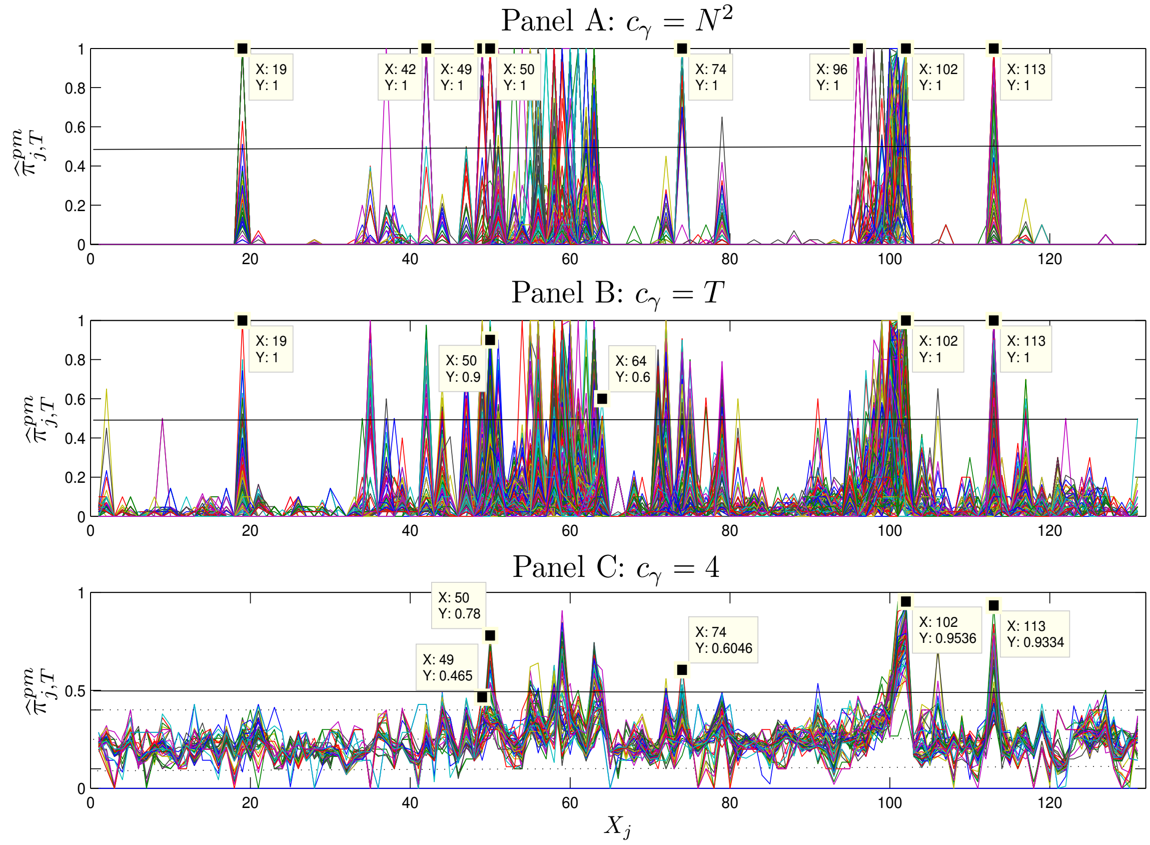
The ridge regression shrinks all coefficients towards zero with more shrinkage on low-variance directions. This means that the ridge will results in many small coefficients. As the shrinkage penalization ν increases so does the number of non-zero coefficients in $\widehat{\beta}^{ridge}$. A high shrinkage parameter ν corresponds to a small tuning parameter c_γ ($c_\gamma \equiv 1/\nu$). This may explain why the highest correlation between the BMA and ridge forecasts occurs when $c_\gamma = 4$ with a 80% explained in sample variance.

The PC regression leaves the r directions with the highest variance alone and discards the remaining $N - r$ directions. The lasso also truncates at zero and results in r large coefficients and sets the remaining $N - r$ to zero. This may explain the similarities of the patterns observed in the the sample correlation between BMA forecasts and those generated by lasso and PCR. In the last column of the first Panel in Table 1, we report the BMA estimate of the model size for the three priors. The results reflect the amount of shrinkage implied by these choices of c_γ. The size of the posterior mean model is decreasing in c_γ with $c_\gamma = N^2$ resulting in the smallest posterior mean estimate of the model size. We observe that the BMA estimate for the model size $\widehat{q}_{pm} = 2.55$ under $c_\gamma = N^2$ and the maximum correlation between BMA and both PCR and lasso forecasts is reached when $r = 3$. We also have notice that under $c_\gamma = T$, $\widehat{q}_{pm} = 7$ and the maximum correlation between BMA forecasts and lasso occurs for $r = 10$ and for BMA and PCR forecasts this number is $r = 3$. Finally for $c_\gamma = 4$, the maximum correlation between BMA and both lasso and PCR forecasts is at $r = 50$ at the same time we have $\widehat{q}_{pm} = 32$.

To examine the relative performance of BMA compared to PCR, we report the MSFE relative to the random walk and the variance (number in parenthesis) of the forecasts relative to the variance of the series to be forecast in Table 2. Under each MSFE row, we report the variance of the forecast relative to the variance of the series. We examine the results for BMA_X which refers to the econometric model (7) where we apply BMA directly to all available predictors in $\mathbf{X}$. In terms of MSFE and over the three sample periods, PCR performs its best when $r = 10$ for industrial production and $r = 5$ for consumer price index. It also outperforms BMA for all the choices of c_γ. However, BMA forecasts tend to have lower variance relative to the forecasts of the series of interest. This observation holds also for the consumer price index forecasts.

Figure 1 and Figure 2 plot of the out-of-sample 12-steps ahead monthly forecasts for industrial production and consumer price index, respectively. These figures One can see the poor performance of all methods in the last subperiod from $1985 - 2002$. We can also see the better performance of principal components regression forecasts in the early months of subperiod $1971 - 1984$, especially forecasting the 1975 recession.

Figure 7: The BMA estimates of the inclusion probabilities $\widehat{\pi}^{pm}_{j,T}$ for $j = 1, \cdots, N$ and $T = T_0 \cdots, T_1$. The $x - axis$ represents the predictors index j, the $y - axis$ the value of the posterior probability of inclusion of X_j in the forecasting model, and each line in the plot represents a different value of T. We also show some of the predictors that appear to be in the median model for several values of T and for the three choices of c_γ.



5 Sensitivity to priors settings

It is worth noting that, we forecast *ip* and *cpi* jointly. That is the Bayesian variable selection is performed for the system of two equations $\mathbf{y}_t \equiv (ip_t, cpi_t)$. Therefore the posterior inference is conditional on $\mathbf{y}$ and $\mathbf{X}$. We examine the mean of the posterior inclusion probabilities $\pi_{j,T} = P(\gamma_j = 1|\mathbf{y}, \mathbf{X})$ as well as its distribution over time.

In terms of local relative performance, Figure 3 and Figure 4 strongly suggest that for the purpose of out-sample forecasting, the choice of the prior on c does not affect the forecast performance of BMA relative to the competing models.

Figure 8: Snapshots of the Median model

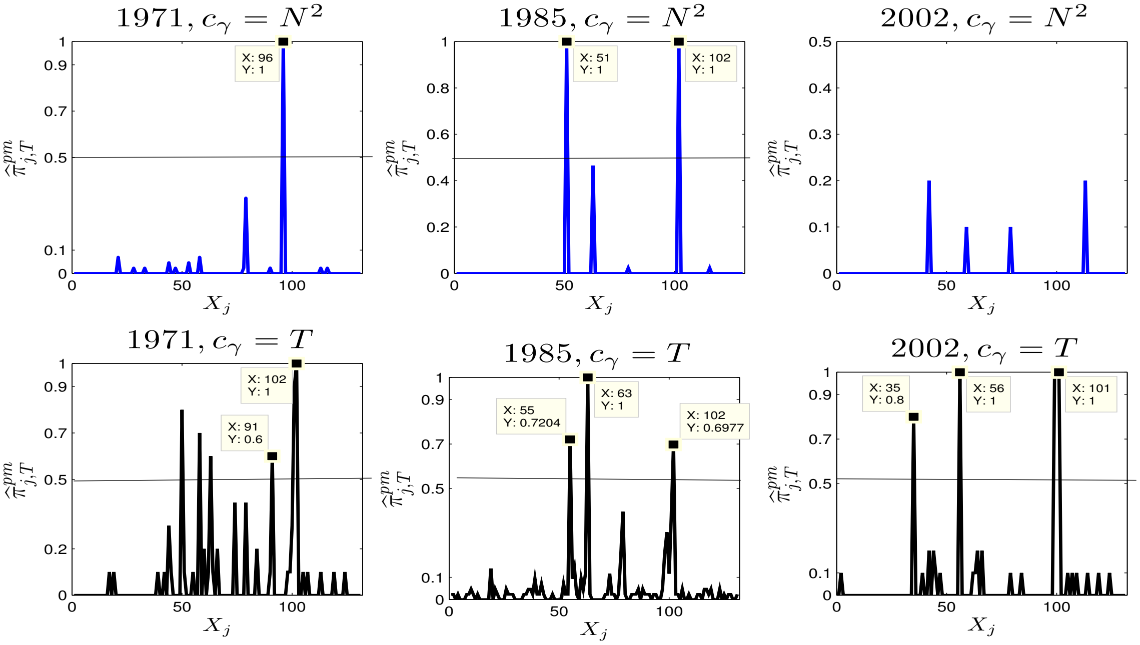
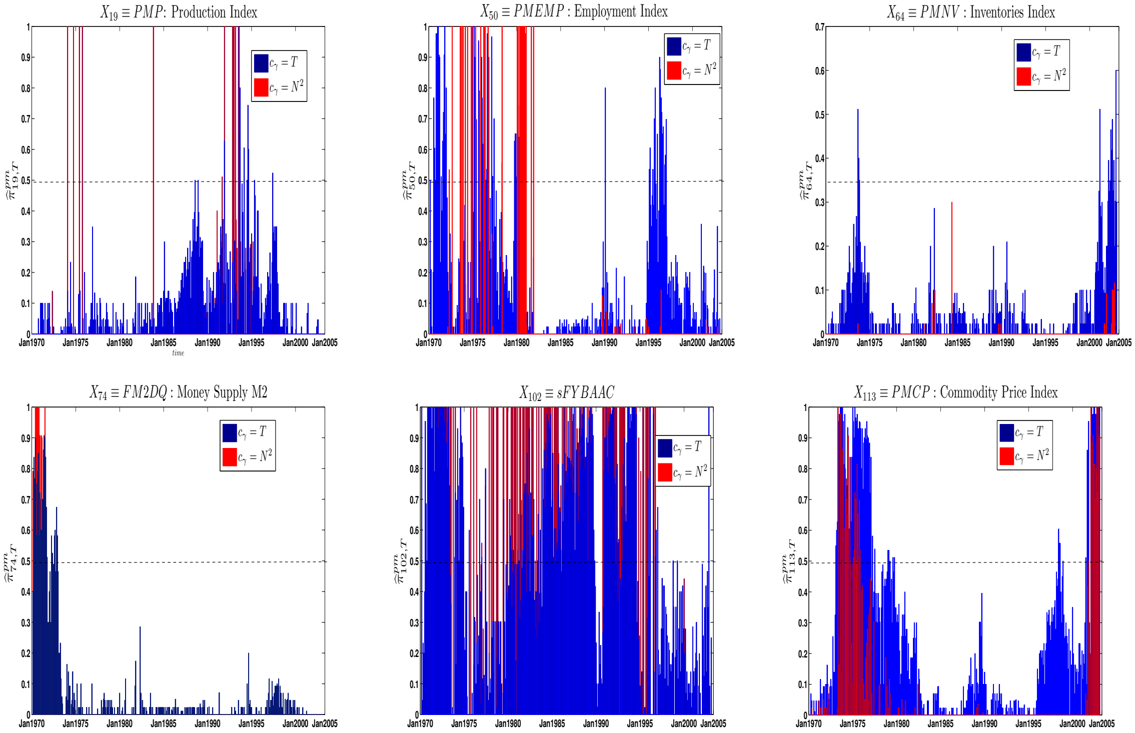


Figure 7 shows the plot of $\widehat{\pi}_{j,T}$ for $T = T_0 : T_1$ under the three choices of tuning parameter c_γ. Each line in the plots corresponds to a different value of T. We note the following patterns. First, in Panel C under the prior $c_\gamma = 4$, while there is a cluster of the probabilities between 0.10 and 0.40, none of the posterior inclusion probabilities reaches the value of 1 with a maximum of 0.96. This reflects the shrinkage effect under $c_\gamma = 4$ where all coefficients are shrunk towards zero resulting in nonzero posterior marginal probabilities for many predictors. To be able to distribute the unit mass over many predictors, the resulting probabilities are in turns very small. Similar behavior was discussed in De Mol et al. (2008) with respect to the Ridge regression coefficients. In contrast, under the risk information prior in Panel A, there seems to be a bimodal behavior. Many predictors appear to have zero posterior inclusion probability and the others have their probabilities equal to one. In Panel B under the unit information $c_\gamma = T$, we see an intermediate behavior with many probabilities equal to zero, others reaching 1 but also a cluster around intermediate values. We conjecture that the risk information prior behaves similar to the lasso in the sense that it forces parsimony with posterior inclusion probabilities that are either zero or one therefore resulting in smaller size models. Second, it appears there are clusters of predictors with high collinearity. By examining Panel A and Panel B, the predictor with posterior inclusion probability one belong to these groups: industrial production ($j = 19$), sectoral ($j = 35 \cdots, 49$), employment index ($j = 50$), housing ($j = 51, \cdots, 60$), Money Supply/Stock ($j = 51, \cdots, 60$) and ($j = 96, \cdots, 102$). Under the risk and unit information prior, at any point in time T the median model will have *at most* one predictor from each of these groups with probability 1. Figure 8 shows a snapshot of the median model at $T = 1971 : 12, 1985 : 12, 2002 : 12$. Under the prior $c_\gamma = 4$, at any time T, a combination of predictors in these clusters will show in the forecasting model, although they will not necessarily appear in the median model. Third, there are some predictors

Figure 9: The time-varying posterior mean estimates of inclusion probabilities for some "important" predictors

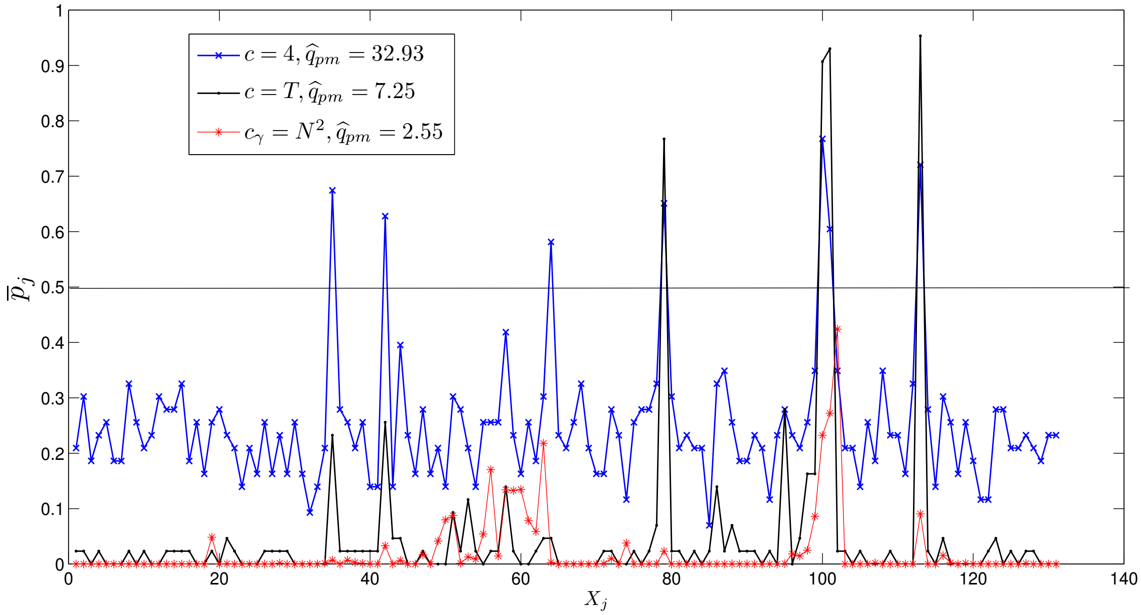


that appear to be "popular" under the three choices of priors. Some of these predictors are examined in Figure 9.

Next, we assess the stability of the forecasting model over time by examining the distribution of $\widehat{\pi}^{pm}_{j,T}$ for those X_j that appear frequently in the median model. Figure 9 shows that the posterior inclusion probability is time-varying. Except for X_{102} which appears to be in the median model for most of the evaluation period, the information value in the other predictors changes over time. The production index turns out to be very informative during the 1990, while the commodity price index plays a role in forcasting both *ip* and *cpi* during the 1970's and 2002.

Figure 10 plots the values of $\overline{\pi}_j$, the time average of the posterior probabilities for the predictors in $\mathbf{X}$ computed using equation (27). Table 4 reports those $\overline{\pi}_j$ that are ≥ 0.5 indicating the *average median* model in the time averaged system. The median model consisting of variables with posterior inclusion probability of at least 0.5 has 7 predictors. Under the unit information prior ($c_\gamma = T$), only 12 predictors have their posterior inclusion probability higher than 10%. The median model consists of only 4 variables and is nested in the median model under $c_\gamma = 4$ with higher

Figure 10: Average (over time) of the posterior mean estimates of inclusion probabilities



inclusion probabilities of most variables in the 90%. Under the risk information prior, the median model is empty with all many variables having their inclusion probability equal to zero. In Table 4, predictor (sFYBAAC) has the highest posterior probability (0.43) under $c_\gamma = N^2$. This predictor does not appear to be in the *average median model* for the other priors.

6 Orthogonal regressors

We compare the out-of-sample performance of BMA using the variables in X as potential predictors and using the principal components of X. We assume that X is generated as follows,

$$X_{jt} = \eta_j' F_t + \epsilon_{jt}, \quad j = 1, ..., N. \tag{33}$$

The $r \times 1$ ($r << N$) vector F_t is a set of common factors driving the dynamics of the cross-section of variables in X. The latent factors F_t are not observed and are replaced with their consistent estimates $\widehat{F}_t$. The factors are estimated by the method of principal components from the panel of data consisting of the X_{jt}, $j = 1, \cdots, N$, $t =, \cdots, T$. The estimated factors $\widehat{F} = \left(\widehat{F}_1, \cdots, \widehat{F}_T,\right)$ is a $T \times N$ matrix consisting of

$$\widehat{F} = X \times v \times D^{-1/2}$$

where v is a matrix of eigenvectors corresponding to the largest $K = min(T, N)$ eigenvalues of $X'X/TN$, and D is a diagonal matrix consisting of the K largest eigenvalues.

Instead of using $\mathbf{X}_t$ as predictors of $\mathbf{y}_t$, in this section we propose the use of $\widehat{F}_t$. Thus we consider the forecasting model:

$$\mathbf{y}_{t+h} = \delta' \widehat{F}_t + \epsilon_{t+h}, \tag{34}$$

In factor analysis, the size of the eigenvalues is related to the amount of information extracted from the explanatory variables and not the dependent variable. It is possible that

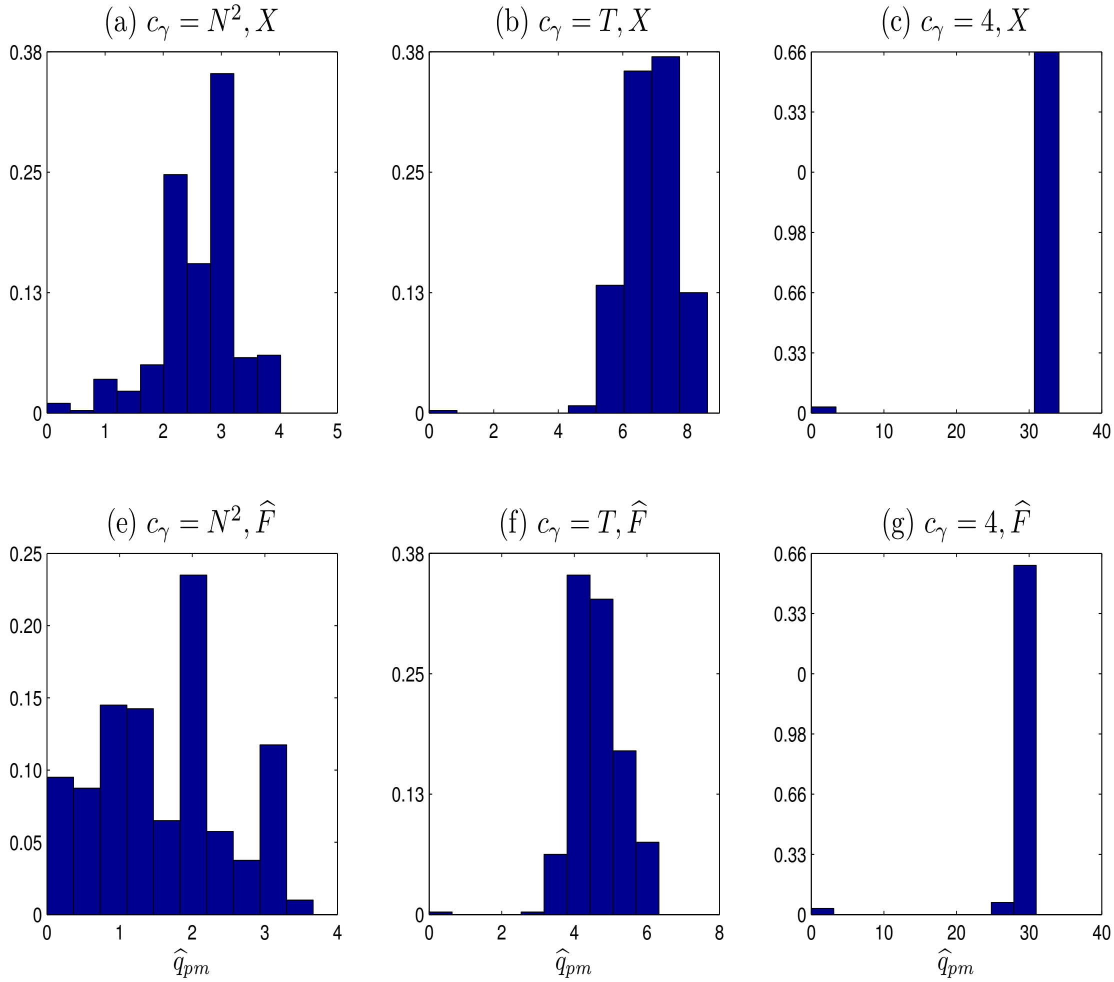
Table 4: Average over time of the BMA estimates of the posterior inclusion probability $\pi_{j,T}$ for X_j in the median average model. The column "lasso" shows that these variables were selected by lasso in forecasting either *cpi*, *ip* or both. The mean of the distribution of the BMA estimate of the size of the forecasting model is denoted $\widehat{\mathrm{E}}\ (\widehat{q}_{pm})$, its standard deviation $std(\widehat{q}_{pm})$ and its median $\widehat{q}_{med}$. We also report the size of the median model (*median*) after (time) averaging the inclusion probabilities.

X_j	$c_\gamma = 4$	$c_\gamma = T$	$c_\gamma = N^2$	lasso
CES006	0.67	-	-	cpi
CES049	0.63	-	-	ip
PMNV	0.58	-	-	cpi
FCLBMC	0.65	0.77	-	cpi
Sfygt10	0.77	0.91	-	ip
sfyaac	0.6	0.93	-	-
PMCP	0.72	0.95	-	ip/cpi
sFYBAAC	*0.35*	*0.03*	*0.43*	
$\widehat{\mathrm{E}}\ (\widehat{q}_{pm})$	32.93	7.25	2.55	
$std(\widehat{q}_{pm})$	4.19	0.84	0.69	
$\widehat{q}_{med}$	32.25	6.90	2.79	
median	7	4	0	

Table 5: Summary statistics for $\widehat{q}_{pm,T}$ under model (7) and model (34). *corr* is the sample correlation coefficient between the two series of $\widehat{q}_{pm,T}$.

	Model with **X**			Model with $\widehat{F}$			
	Mean	std	Median	Mean	std	Median	*corr*
$c_\gamma = 4$	31.8540	4.1940	32.2542	29.5650	3.9182	30.2454	0.9815
$c_\gamma = T$	6.8845	0.8393	6.9000	4.6169	0.6737	4.5255	0.4381
$c_\gamma = N^2$	2.6174	0.6863	2.7910	1.6036	0.8867	1.5898	0.3561

Figure 11: Comparison of the distribution of the BMA estimates of the size of the posterior model, $\widehat{q}_{pm,T}$, using the predictors $\mathbf{X}$ in panels (a), (b) and (c) and their principal components $\widehat{F}$ in panels (e), (f) and (g).



(a) $c_\gamma = N^2, X$

(b) $c_\gamma = T, X$

(c) $c_\gamma = 4, X$

(e) $c_\gamma = N^2, \widehat{F}$

(f) $c_\gamma = T, \widehat{F}$

(g) $c_\gamma = 4, \widehat{F}$

some factors associated with large eigenvalues have no explanatory power while some with small eigenvalues do have explanatory power for the dependent variable.

Our approach is to apply the same methodology described above to perform Bayesian variable selection over the model space defined by the estimated orthogonal factors $\widehat{F}_t$.

Figure 11 compares the distribution of the BMA estimate of the size of the forecasting model $\widehat{q}_{pm,T}$ under model (7) and model (34). Table 5 reports the mean, standard deviation and median of these two series of estimates $\widehat{q}_{pm,T}$. There are strong similarities between the two distributions and strong correlation especially under the prior $c_\gamma = 4$. More important, there is evidence that principal components beyond the first ones are found to be *informative* in forecasting inflation and industrial production. This can be seen clearly in Figure **??** where we show the distribution over time of the posterior inclusion probabilities $\widehat{\pi}^{pm}_{j,T}$ for all $K = min\{T, N\}$ estimated factors $\widehat{F}_j$. Our discussion for the case with $\mathbf{X}$ about the three priors carries over the case with $\widehat{F}$. The unit information and risk information priors both favor few predictors with probabilities equal to one while assigning zero posterior probability on the remaining principal components. Therefore resulting in smaller size posterior models. Even in the case of principal components, the prior $c_\gamma = 4$ shrinks all the probabilities resulting in cluster between 20% and 40%. The interesting fact is that principal components up to $j = 27$ can still appear in the median model. These inclusion probabilities are time varying as can be seen in the examples of Figure 14 where we report $\widehat{\pi}_{j,T}$ for the first eight principal components.

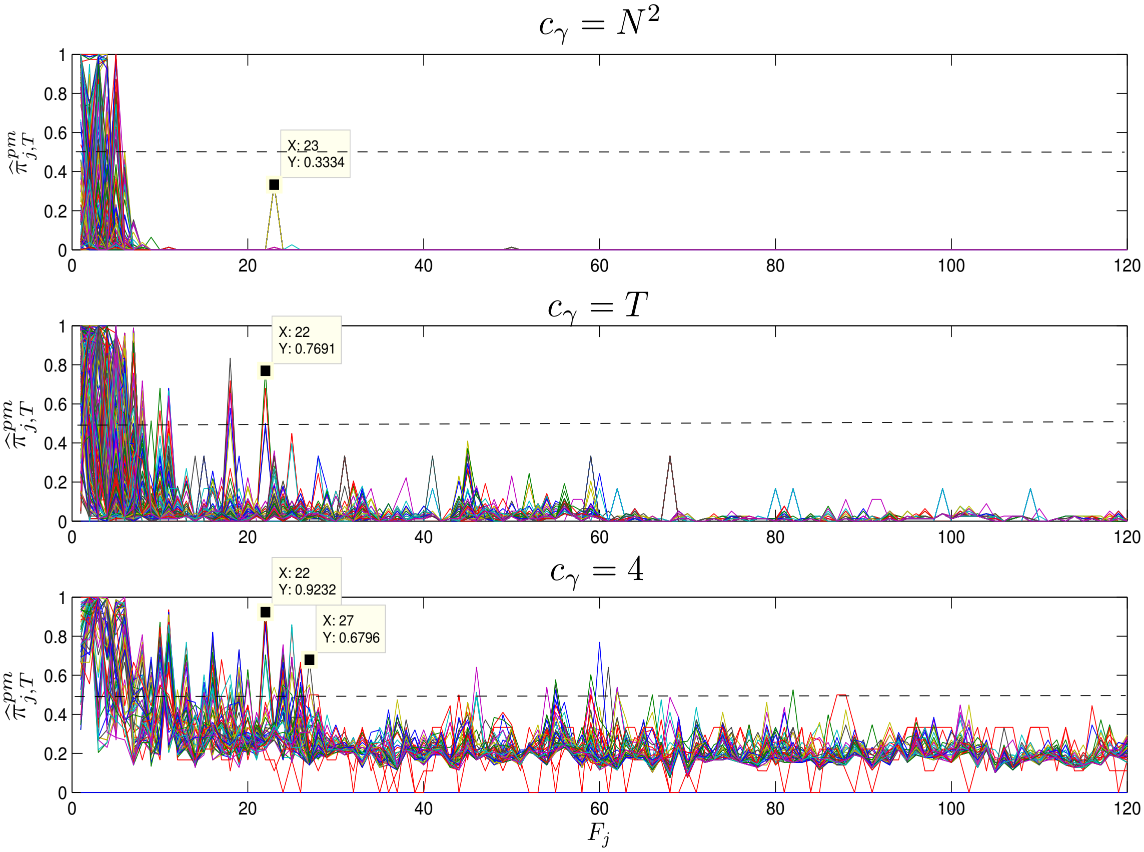
The results indicate that, with the exception of the case of the forecasting consumer price index in the post 1985, applying BMA to orthogonalized predictors does not result in better out-of-sample forecast performance. Figure 13 plots the out-of-sample forecasts for industrial production using $\mathbf{X}$ (denoted BMA) and using $\widehat{F}$ (denoted BMA_F). There is correlation between the two forecasts in most of the early period of the evaluation sample. However during $1985 - 2002$, forecasts based on the model with $\widehat{F}$ as possible predictors performs extremely bad and is very close to the random walk forecasts. Under $c_\gamma = T$, the correlation between the forecasts is about 0.64 for industrial production and 0.58 for inflation. The forecasts using $\widehat{F}$ tend to be more volatile resulting in wider probability intervals compared to those based on $\mathbf{X}$.

An interesting result that is worth highlighting is that the posterior distribution of the model size when BMA is applied to the estimated principal components $\widehat{F}_j, (j = 1, \cdots, N)$, has moments that are similar to those we get when BMA is applied to the predictor variables $\mathbf{X}_j$. Table 5 shows summary statistics of the distribution of $\widehat{q}_{pm,T}$, the posterior mean of the average size of the forecasting model. Noteworthy to note the high correlation between the two distribution (.98). The other priors on c_γ generate distributions that are less correlated to those under $\mathbf{X}$ but the first and second moments are very similar.

Table 6: Correlation of BMA forecasts with principal component forecasts

Industrial Production										
				Number of principal components						
		BMA_X	BMA_F	1	3	5	10	25	50	75
$c_\gamma = T$	BMA_X	1	0.64	0.57	0.73	0.67	0.68	0.69	0.65	0.54
	BMA_F	0.64	1	0.47	0.65	0.65	0.65	0.63	0.61	0.45
Consumer Price Index										
				Number of principal components						
		BMA_X	BMA_F	1	3	5	10	25	50	75
$c_\gamma = T$	BMA_X	1	0.58	0.51	0.56	0.54	0.51	0.60	0.54	0.32
	BMA_F	0.58	1	0.55	0.75	0.74	0.76	0.67	0.39	0.12

Figure 12: The Median Model over time for BMA with principal components as predictors.



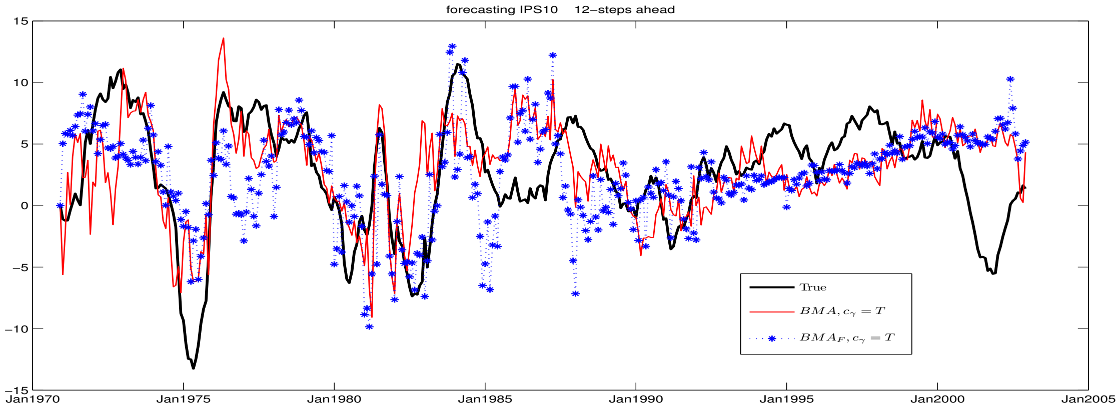
7 Conclusion

This paper analyses the performance of principal components regression and Bayesian model averaging as two competing approaches to forecasting with large datasets. We consider real time forecasting by analysing out-of-sample performance.

We find that in terms of average performance measure such as the root mean-squared-forecast error, PC forecasts are marginally better than BMA although the forecast are more volatile. However, this edge in global relative performance of PC hides considerable changes in the relative local forecasting performance. We find that BMA performance surprisingly good for most of the period pre 1990's. During the 90's, both PC and BMA fail to outperform the naive random walk. The marginal benefit of PC comes from the improved relative performance in the years post 2002. Another important point that this paper highlights is that BMA forecasts series does remarkably well in matching the mean of the actual series with low or almost zero bias. The bias proportion in the PC forecasts series is extremly high especially for the consumer price index. In terms of mean and volatility of the predicted series, BMA generally performs better in matching these moments.

In this paper we also considered applying BMA to orthogonalized predictors in the form of principal components of the predictors variables. Surprisingly, in terms of global forecast performance, this strategy produces out-of-sample forecast series that performed worse than both PC and BMA (applied to the original predictors). The posterior distribution of the average size of the forecasting model is practically the same as the one produced applying BMA to the actual predictors.

Figure 13: IP forecasts based on BMA applied to the principal components of X.



The posterior distribution of the forecasting model generated by BMA indicates that the 'best' forecasting model is time varying as well as the averaging weights. This further highlights findings in the literature (Giacomini and Rossi (2010)) about the existence of instabilities in the forecasting environment.

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 135–172.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 307–317.

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *The Annals of Statistics 32*(3), 870–897.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operation Research Qarterly 20*, 451–468.

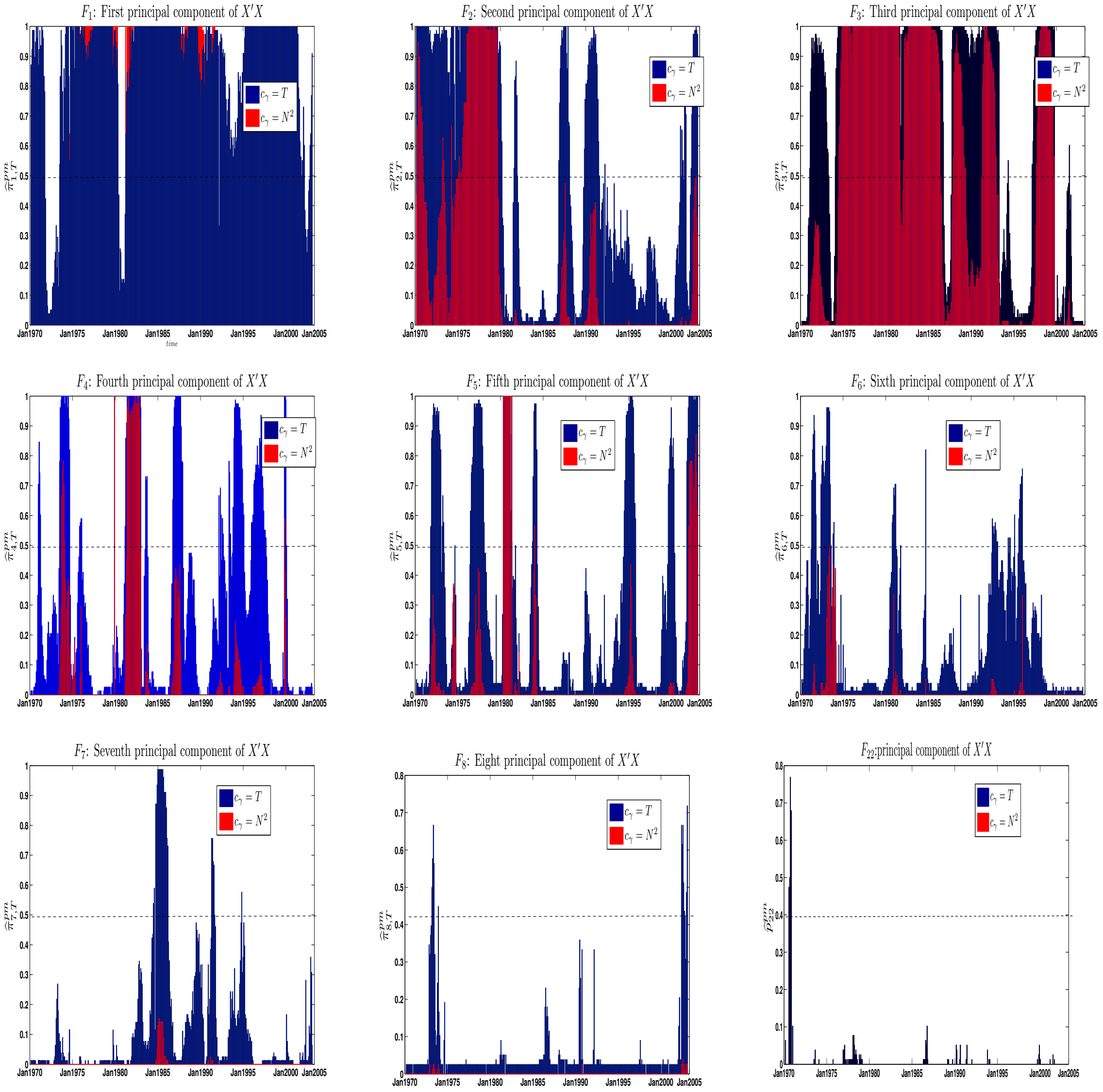
Brown, P. J., T. Fearn, and M. Vanucci (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika 86*(3), 635–6487.

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

Chauvet, M. and S. Potter (2012). Forecasting output. *The Handbook of Economic Forecasting 2*.

Chipman, H., E. I. George, and R. E. McCulloch (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes- Monograph Series 38*, 65–134.

Figure 14: The time-varying posterior mean estimates of inclusion probabilities for the first principal components F_j of $\mathbf{X}'\mathbf{X}$



Clemen, R. T. and R. L. Winkler (1986). Combining economic forecasts. *Journal of Business & Economic Statistics 4*(1), 39–46.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. *Bayesian Statistics 6*, 157–185.

Danilov, D. and J. R. Magnus (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics 122*, 27–46.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*, 318–328.

Diebold, F. and J. Lopez (1996). Forecast evaluation and combination. *In: Maddala, Rao (Eds.), Handbook of Statistics Elsevier, Amsterdam.*

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*, 425–456.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regresson. *Annals of Statistics 32*(2), 407–499.

Fernandez, C., E. Ley, and M. Steel (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics 100*, 381–427.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). Opening the black box: Structural factor models with large cross-sections. *Unpublished manuscript, Universite Libre de Bruxelles*.

George, E. I. and D. P. Foster (2000). Calibration and empirical bayes variable selection. *Biometrika 87*(4), 731–747.

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*, 595–620.

Giannone, D., L. Reichlin, and L. Sala (2004). Monetary policy in real time. *In Market Gertler and Kenneth Rogoff editors, NBER Macroeconomics*, 161–200.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica 75*(4), 1175–1189.

Hendry, D. F. and M. P. Clements (2002). Pooling of forecasts. *Econometrics Journal 5*, 1–26.

Hjort, N. L. and G. Claeskens (2003). Frequentist model average estimators. *Journal of the American Statistical Associations 98*, 879–899.

Jacobson, T. and S. Karlsson (2004). Finding good predictors for inflation: a bayesian model averaging approach. *Journal of Forecasting 23*(7), 479–496.

Kohn, R., M. Smith, and D. Chan (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing 11*, 313–322.

Komunjer, I. and M. T. Owyang (2011). Multivariate forecast evaluation and rationality testing. *The Review of Economics and Statistics Forthcoming.*

Koop, G. and S. Potter (2004). Forecasting in dynamic factor models using Bayesian model averaging. *Econometrics Journal 7*(2), 550–565.

Leeb, H. and B. M. P`otcher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics 342*, 2554–2591.

Madigan, D. and A. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association 89*, 1535–1546.

Ouysse, R. and R. Kohn (2009). Bayesian variable selection and model averaging in the arbitrage pricing theory. *Computational Statistics and Data Analysis forthcoming.*

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*, 1790–191.

Smith, A. and G. Roberts (1993). Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B 55*, 3–24.

Stock, J. H. and M. . Watson (2006). Forecasting with many predictors. *In Handbook of Economic Forecasting 1*, 551–554.

Stock, J. H. and M. W. Watson (1998). Diffusion indexes. *NBER Working Paper 6702.*

Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.

Stock, J. H. and M. W. Watson (2005). An empirical comparison of methods for forecasting using many predictors. *Manuscript. Princeton University*.

Theil, H. (1967). Economics and information theory. *Amsterdam: North-Holland*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Seris B 58*, 267–288.

Wright, J. H. (2009). Forecasting u.s. in ation by Bayesian model averaging. *Journal of Forecasting 28*, 131–144.

Zellner, A. (1986). Further results on Bayesian minimum expected loss (MELO) estimates and posterior distributions for structural coefficients. *In Slottje, D., eds.,* Advances in Econometrics *5*, 171–182.