Comparison of Bayesian Moving Average and Principal Component Forecasts for Large Dimensional Factor Models

Rachida Ouysse [a]

First draft: September 22, 2010
This version: April 6, 2011

[a]*The Australian School of Business, The University of New South Wales UNSW, Sydney NSW 2052*
Email: rouysse@unsw.edu.au

Abstract: The growing availability of financial and macroeconomic data sets including a large number of time series (hence the high dimensionality) calls for econometric methods providing a convenient and parsimonious representation of the covariance structure both in the time and the cross-sectional dimensions. Currently, dynamic factor models constitute the dominant framework across many disciplines for formal compression of information. Recent econometric research has produced a rich body of theory for the estimation of these models and their subsequent use for forecasting and for the estimation of structural economic models.

To overcome the challenges of dimensionality, many forecast approaches proceed by somehow reducing the number of predictors. Principal component regression (PCR) approach proposes computing forecasts as projection on the first few principal components of the predictors. Bayesian model averaging (BMA) approach combines forecasts to extract information from different possible relationships between the predicted variable and the predictor variables. These two literature 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.

This paper provides empirical evidence for connecting these two seemingly different approaches to forecasting. We study the performance of BMA as a forecasting method based on large panels of time series as an alternative to 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. The results can also be driven by the nature of the macroeconomic data which is characterized by collinearity and that the variable selection is sensitive to minor perturbations of the data. The empirical results serve as a preliminary guide to understanding the behavior of BMA under double asymptotics, i.e. when the cross-section and the sample size become large.

Keywords: Bayesian variable selection, shrinkage regression, principal components analysis, factor models, forecasting.

1 Introduction

This study provides empirical evidence for connecting these two seemingly different approaches to forecasting. With the exception of De Mol et al. [2008] who compare the forecasts performance of PCR and Bayesian shrinkage, little is known about the links between BMA and PCR forecasts. De Mol et al. [2008] study the empirical and theoretical properties of Bayesian shrinkage and Ridge regression forecasts and compared them with PCR forecasts. They find that the two methods produce forecasts which are highly correlated with similar out-of-sample performance. De Mol et al. [2008] are the first to 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. These findings are consistent with the convergence result shown by Ouysse [2006] in the context of classical analysis of factor models.

Using the notation in De Mol et al. [2008], consider the $(n \times 1)$ vector of covariance stationary processes $Z_t = (z_{1t}, \cdot\cdot\cdot, z_{nt})'$ 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, $\mathbf{y}_{t+h|t} = proj\{\mathbf{y}_{t+h}|I_t\}$, where $I_t = span\{Z_{t-s}, s = 0, 1, 2, \cdots\}$ is a potentially large information set, and $\mathbf{y}_{t+h} = (y_{1,t+h}, \cdots, y_{m,t+h})$ is an $m-$vector of filtered versions of z_{it}, where for specific $i = 1 \cdot\cdot\cdot, n$ and $1 \leq m \leq n$, $y_{j,t+h} = f_{j,h}(L)z_{i,t+h}$ and L is the lag operator defined as $L^l z_t = z_{t-l}$ for any integer l.

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 $X_t = (Z_t', \cdot\cdot\cdot, Z_{t-p}')$ for each target series j, $j = 1, \cdots, m$. Given a sample of size T, let $\mathbf{X} = (X_{p+1}, \cdot\cdot\cdot, X_{T-h})'$ 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 $\widehat{y}^{LS}_{j,T+h|T} = \mathbf{X}'\widehat{\beta}^{LS}$, where $\widehat{\beta}^{LS}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y_j$, $j = 1, \cdots, m$.

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. There are three strands of the literature on forecasting using large datasets. The first uses factor models and principal components regression (PCR). The second shrinks 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 which combines forecasts from an ensemble of models. In this study, we compare the out-of-sample performance of PCR and BMA based forecasts. We find that these are highly correlated with marginal differences suggesting that BMA and PCR may in fact be two sides of the same coin.

2 Principal component regression

We consider forecasting situation in which both N and T are large, hence the double (N, T) asymptotics with 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 modeling 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 (PCA) 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 [2002] 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 assumptions 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}). The factors F_t are unobserved and the number of common factors r is also unknown. Principal components regression (PCR) computes the forecasts as a projection on the first few principal components. Let $\widehat{F}_t$ be the $T \times r$ matrix of the first r principal components of the predictors $\mathbf{X}$ and let $I_t^f = span\{\widehat{f}_{1t}, \cdot\cdot\cdot, \widehat{f}_{rt}\}$ with $r \ll N$ be a parsimonious representation of the information set 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: $S_x V = VD$ 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:

$$y_{j,T+h|T}^{PC} = proj\{y_{j,T+h}|I_T^f\}. \tag{3}$$

Once the factors are estimated via PCA, the projection is computed by OLS treating the factors as observed:

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

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

2.1 Shrinkage regression

Ridge regression and the lasso are classical approaches to shrinkage regression defined as:

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

for some penalization parameter $\lambda \geq 0$. Choosing $\kappa = 2$ yields ridge regression where $\widehat{\beta}_j^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda I_N)^{-1} \mathbf{X} y_j$. Choosing $\kappa = 1$ yields the lasso which has no closed form solution but the entire path of λ can be obtained using the LARS algorithm. 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 $\beta_j|\sigma_\epsilon^2 \sim \mathcal{N}(0, \sigma_\epsilon^2\lambda)$; for the lasso it is an independent identically distributed Laplace (double exponential) $p(\beta_{j,k}|\sigma_\epsilon^2) = \frac{\lambda}{2\sigma_\epsilon} e^{-\lambda|\beta_{j,k}|/\sigma_\epsilon}$.

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 optimal forecast obtained if factors are observed. 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.

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.

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:

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

A commonly used prior for γ is

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

with π prespecified. We follow Fernandez et al. [2001] and choose $\pi = 0.5$ implying that $p(\gamma) = 2^{-N}$. Using a Normal inverse-Wishart conjugate prior, we implement Bayesian variable selection by specifying a g-prior for $\beta|\Sigma$ as $N(0, c\Sigma \otimes (\mathbf{X}'\mathbf{X})^{-1})$. 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. 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 $c = N^2$. A conjugate g-prior with fixed $c \cong 4$ corresponds asymptotically to Akaike's AIC. 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 mean $\widetilde{\beta}_\gamma$ of the posterior density $p(\beta|\mathbf{y}, \Sigma, \gamma)$ is $\widetilde{\beta}_\gamma = \eta_\gamma \widehat{\beta}_\gamma$ 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.

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

Forecast period $1970:12$ **to** $2002:12$									
Correlation of forecasts: LASSO with BMA									
	Number of non zero coefficients								
	1	3	5	10	25	50	75		$\widehat{\mathrm{E}}\ (\widehat{q}_{pm})$
$c_\gamma = T$	0.43	0.74	0.80	0.86	0.85	0.78	0.61		7.25
$c_\gamma = N^2$	0.50	0.82	0.85	0.85	0.78	0.69	0.51		2.55
$c_\gamma = 4$	0.49	0.75	0.80	0.87	0.91	0.91	0.80		32
Correlation of forecasts: 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
$c_\gamma = T$	0.65	0.77	0.81	0.82	0.81	0.79	0.74	0.64	0.41
$c_\gamma = N^2$	0.57	0.73	0.79	0.83	0.83	0.82	0.78	0.69	0.46
$c_\gamma = 4$	0.84	0.90	0.89	0.87	0.85	0.82	0.77	0.69	0.50
Correlation of forecasts: PC with BMA									
	Number of principal components, r								
	1	3	5	10	25	50	75		
$c_\gamma = T$	0.21	0.72	0.77	0.79	0.79	0.73	0.61		
$c_\gamma = N^2$	0.26	0.77	0.82	0.83	0.80	0.66	0.50		
$c_\gamma = 4$	0.16	0.69	0.72	0.76	0.79	0.82	0.71		

Note that in the case the target variables in $\mathbf{y}$ are predicted equation by equation and the prior on β_j has a prior $\mathcal{N}(0, \sigma^2_{\epsilon_j} c I_N)$ with data independent covariance, 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 $c_\gamma \equiv \frac{\sigma^2_{\beta_j}}{\sigma^2_{\epsilon_j}}$, see 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.

In BMA the posterior distributions of quantities of interest are obtained as mixtures of the model-specific distributions weighted by the posterior model probabilities. The BMA estimate of the posterior predictive density of $\mathbf{y}_{t+h}$, conditional on $\mathbf{y}$ and $\mathbf{X}$ (the information at time T) is:

$$p(\mathbf{y}_{T+h} | \mathbf{y}, \mathbf{X}) \quad = \quad \sum_\gamma p(\mathbf{y}_{T+h} | \mathbf{y}, \mathbf{X}, \gamma) p(\gamma | \mathbf{y}, \mathbf{X}). \tag{10}$$

The BMA forecast for $\mathbf{y}_{t+h}$, defined as the expected value of the density in (10), is

$$\widehat{\mathbf{y}}^{BMA}_{T+h|T} \quad = \quad \sum_\gamma (I_m \otimes \mathbf{X}_\gamma) \widetilde{\beta}_\gamma p(\gamma | \mathbf{y}, \mathbf{X}). \tag{11}$$

Implementation of (11) 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 $p(\gamma|\mathbf{y})$; $\gamma^{(j)}, j = 1, ..., M$. The quantity in (11) is therefore approximated using

$$\widehat{\mathbf{y}}^{pm}_{T+h|T} \quad = \quad \frac{1}{M} \sum_{j=1}^{M} (I_m \otimes \mathbf{X}_{\gamma^{(j)}}) \widetilde{\beta}_{\gamma^{(j)}}, \tag{12}$$

where $\gamma^{(j)}$ is the posterior model in the j^{th} MCMC iteration and M is the number of MCMC iterations.

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]. The total number of predictors $N = 131$ in $\mathbf{X}$ includes real variables such as sectoral industrial production, employment and hours worked; nominal variables such as consumer and price indices, 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, etc) and first differences for variables already expressed in rates (unemployment rate, capacity utilization, etc).

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

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

$IP_T = 100 \log IP_t$ is the rescaled log 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. In this section we compare the performance of BMA forecasts to those based on principal components and shrinkage (ridge and lasso) regression. 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.

The results in Table 1 suggest the following. 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 non zero 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 $c_\gamma = 4$.

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 when the number of non zero coefficients is 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 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

Table 2: Comparison of principal component and Bayesian model averaging forecasts.

Industrial Production							
	Bayesian model averaging				Principal Component		
	BMA				r		
	$c_\gamma = T$	$c_\gamma = 4$	$\widehat{c}_\gamma^{EB}$	$c_\gamma = N^2$	5	10	25
MSFE $1971 - 2002$	0.8779	0.8635	0.8592	0.8700	0.56	0.54	0.65
	(0.66)	*(0.51)*	*(0.49)*	*(0.50)*	*(0.97)*	*(1.28)*	
MSFE $1971 - 1984$	0.5645	0.6997	0.6882	0.7039	0.35	0.34	0.46
	(0.54)	*(0.47)*	*(0.44)*	*(0.45)*	*(0.93)*	*(1.11)*	*(1.43)*
MSFE $1985 - 2002$	1.8071	1.3490	1.3664	1.3625	1.16	1.13	1.21
	(1.01)	*(0.62)*	*(0.63)*	*(0.62)*	*(0.33)*	*(0.51)*	*(0.79)*
Consumer Price Index							
	Bayesian model averaging				Principal Component		
	BMA				r		
	$c_\gamma = T$	$c_\gamma = 4$	$\widehat{c}_\gamma^{EB}$	$c_\gamma = N^2$	5	10	25
MSFE $1971 - 2002$	0.7861	0.7761	0.8045	0.7777	0.57	0.69	0.83
	(0.50)	*(0.52)*	*(0.53)*	*(0.52)*	*(0.61)*	*(0.63)*	*(0.69)*
MSFE $1971 - 1984$	0.6773	0.6789	0.7137	0.6839	0.39	0.48	0.56
	(0.49)	*(0.49)*	*(0.50)*	*(0.50)*	*(0.57)*	*(0.57)*	*(0.60)*
MSFE $1985 - 2002$	1.2970	1.2327	1.2308	1.2179	1.43	1.71	2.11
	(0.53)	*(0.61)*	*(0.61)*	*(0.59)*	*(0.73)*	*(0.83)*	*(0.95)*

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.

5 Conclusions and Recommendations

To overcome the challenges of dimensionality in forecasting with large number of predictors, PCA and BMA stand out as the most popular methods in the recent literature. This study compares these seemingly unrelated approaches in an empirical application. The results are promising and suggest that for the purpose of forecasting, the two approaches are capturing the same information from the data. The out-of-sample forecasts are highly correlated and the two methods are relatively similar in terms of mean squared forecast errors. These results are purely empirical and provide a motivation to establishing the theoretical foundations that link the two approaches in the forecasting framework.

References

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.

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

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

Ouysse, R. (2006). Consistent variable selection in large panels when factors are observable. *Journal of Multivariate Analysis 97*, 946–984.

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

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