International Journal of Forecasting 29 (2013) 43–59

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Hierarchical shrinkage priors for dynamic regressions with many predictors

Dimitris Korobilis

University of Glasgow, United Kingdom

ARTICLE INFO

Keywords:
Forecasting
Shrinkage
Factor model
Variable selection
Bayesian lasso

ABSTRACT

This paper examines the properties of Bayes shrinkage estimators for dynamic regressions that are based on hierarchical versions of the typical normal prior. Various popular penalized least squares estimators for shrinkage and selection in regression models can be recovered using a single hierarchical Bayes formulation. Using 129 US macroeconomic quarterly variables for the period 1959–2010, I extensively evaluate the forecasting properties of Bayesian shrinkage in macroeconomic forecasting with many predictors. The results show that, for particular data series, hierarchical shrinkage dominates factor model forecasts, and hence serves as a valuable addition to the existing methods for handling large dimensional data.

1. Introduction

Over the past 20 years there has been a great deal of work on modeling and forecasting using large macroeconomic data sets. Although the focuses of the models may differ between applications, the basic modeling structure involves one or more variables of interest and a huge number of possible explanatory variables. For many years, such problems have been being addressed in the economics literature using simple static or dynamic factor models (Stock & Watson, 2002). The underlying reason why factor models have been used so extensively in modeling macroeconomic time series is the tractability of factor estimates through the use of principal components.

Nevertheless, in most cases, efficiency in using the information in many variables to increase the forecast accuracy is far more important than tractability. That is why numerous recent studies have examined the potential to use intensive statistical algorithms that can either shrink or completely remove irrelevant predictors when forecasting using dynamic regression models with large macroeconomic data sets. For instance, Inoue and Kilian (2008) make an extensive comparison of the capability of the bagging algorithm to shrink the space of predictors when forecasting U.S. inflation. De Mol, Giannone, and Reichlin (2008) show that forecasts of inflation and industrial production using the least absolute shrinkage and selection operator (lasso) of Tibshirani (1996) are at least as favorable as principal components (factor model) forecasts. Other statistical methods have also been used in economics for handling large information sets, including the boosting algorithm (Bai & Ng, 2007), empirical Bayes shrinkage (Litterman, 1979), Bayesian model averaging (Koop & Potter, 2004), and dynamic model averaging (Koop & Korobilis, 2012).

The purpose of this article is to contribute to this expanding body of literature by comparing the forecasting performances of hierarchical Bayes priors in dynamic regressions with many predictors. In particular, I focus on five prior choices which result in regularized estimators[1] that are Bayesian equivalents of frequentist shrinkage estimators, such as the lasso (Tibshirani, 1996), the fused lasso (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005), and the elastic net (Zou & Hastie, 2005). Due to the small time series dimension of macroeconomic data (usually quarterly or monthly), noninformative choices that allow

E-mail address: dikorombilis@yahoo.gr.

[1] That is, estimators that can help prevent overfitting.

0169-2070/$ – see front matter © 2012 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2012.05.006

prior variances of regression coefficients to be very large do not provide any shrinkage, and hence, most of the time they are not empirically satisfactory. Empirical Bayes priors can be used to provide shrinkage in large data sets (see Litterman, 1979); nonetheless, these priors become so subjective that they can hardly appeal to the inexperienced non-Bayesian researcher. Hierarchical Bayesian inference can be used instead to estimate unknown prior parameters from the data, by allowing the prior variances of regression coefficients to have a hyperprior distribution of their own. That way, hierarchical priors can both provide desirable shrinkage in large data sets, and also provide a basis for a more objective Bayesian analysis by potentially allowing noninformative priors to be used at the hyperprior level.

Hierarchical shrinkage priors are increasing in popularity for statistical applications involving high dimensional and correlated genetic data sets (MacLehose & Dunson, 2010; Yi & Xu, 2008). The development in this area is so vast that Fahrmeir, Kneib, and Konrath (2010), Griffin and Brown (2010) and Kyung, Gill, Ghoshz, and Casella (2010) recently provided flexible unified representations of Bayesian hierarchical priors, while at the same time evaluating their shrinkage performance. Evidently, this article follows a similar route; however, the focus is on forecasting macroeconomic time series using large data sets. For example, the question of whether hierarchical shrinkage works well for the massive genetics data or not tells us little about how it could be useful for an applied economist who typically monitors a few hundred macroeconomic time series to forecast, say, inflation, unemployment and the interest rate. One exception in the economics literature is the recent paper by Giannone, Lenza, and Primiceri (2012), which examines the properties of forecasting using hierarchical priors on the Minnesota shrinkage coefficient in vector autoregressive models (VARs).

Therefore, by using univariate predictive regressions on 129 quarterly macroeconomic and financial time series for the U.S. (where I sequentially use one series as the dependent variable and the remaining 128 as the large set of predictors which are subject to shrinkage), I show that Bayesian hierarchical shrinkage can be quite beneficial in forecasting, compared to both simple AR forecasts and principal components forecasts coming from a factor regression model. In order to provide a basis for reference for future work in this area, my default choice is noninformative priors for the parameters that control regularization/shrinkage. Although they are not calibrated optimally based on the information in the data, noninformative priors on the regularization parameters are shown to actually be efficient in providing shrinkage sufficient to avoid overfitting and erroneous forecasts. In spite of using these noninformative priors as a benchmark, an additional sensitivity analysis of the Bayesian lasso shrinkage for forecasting the gross domestic product (one of the most important macroeconomic time series) concludes this paper. In this case, I contrast the noninformative prior on the regularization parameter(s) with some informative values (selected "subjectively"), as well as with a semi-automatic method of estimating the regularization parameters based on marginal maximum likelihood.

The paper is organized as follows. Section 2 describes the modeling framework and the estimation methodology. The general dynamic regression problem with many predictors is introduced, together with a unified shrinkage representation of Bayes estimators, and specific cases are analyzed in detail. Section 3 describes the forecast methodology and subsequently reports the results from the out-of-sample exercise for five special shrinkage estimators applied on 129 series. This section continues with a sensitivity analysis on the lasso prior. Section 4 concludes and provides an assessment of the empirical value of hierarchical shrinkage priors.

2. Bayes shrinkage formulations for dynamic regressions

In this paper I consider univariate forecasting models of the form

$$y_{t+h} = \boldsymbol{\alpha}'\boldsymbol{z}_t + \boldsymbol{\beta}'\boldsymbol{x}_t + \varepsilon_{t+h}, \tag{1}$$

where ε_{t+h} is an error term distributed as $\varepsilon_t \overset{iid}{\sim} N\left(0, \sigma^2\right)$, for $t = 1, \ldots, T$. In this type of regression, y_{t+h} is the h-quarter-ahead value of the variable of interest, $\boldsymbol{z}_t$ is the $q \times 1$ vector of unrestricted predictors which are always included in the forecasting model, such as the intercept, dummy variables,[2] and lags of the dependent variables, and $\boldsymbol{x}_t$ is the $p \times 1$ vector of exogenous predictors whose dimension we would like to shrink.

The number of coefficients $\boldsymbol{\beta}$ of the predictor variables $\boldsymbol{x}_t$ might be too large relative to the number of observations. In this case, shrinkage is necessary for two reasons. First, even when all predictors are relevant and the full model with all p predictors included is the correct (unbiased) model, we can almost always find a biased model with a lower in-sample mean square error (MSE). By shrinking some of the coefficients in the full model, and hence, introducing some bias, we can achieve a much lower variance of the coefficient estimates and lower the MSE (this is called the "variance-bias trade-off"). Second, heavily parametrized models tend to be overfitted in-sample, and provide very poor out-of-sample fits. By introducing some sort of penalty for very complex models through shrinkage and focusing on a model with a few useful predictors, we can achieve parsimony and enhance the economic interpretation of our results.

The unrestricted vector of coefficients $\boldsymbol{\alpha}$ and the variance σ^2 can be integrated out using the noninformative priors $\pi\left(\boldsymbol{\alpha}\right) \propto 1$ and $\pi\left(\sigma^2\right) \propto 1/\sigma^2$ respectively. This allows a closer focus on the regression coefficient vector $\boldsymbol{\beta}$, which has individual elements $\beta_j, j = 1, \ldots, p$. In addition, in the remainder of this paper all exogenous predictors, $\boldsymbol{x}_t$, are standardized to have a zero mean and variance 1, unless stated otherwise. This transformation is typical in model selection and shrinkage problems, where we need to identify only the good predictors among a large

[2] When predictors are categorical, as is the case with dummy variables, the lasso solution depends on the way in which these predictors are encoded (for instance 0/1/2/3 instead of 1/2/3/4). In this case, the group lasso algorithm of Yuan and Lin (2006) provides a solution to this issue.

set of exogenous variables. The variables in $\boldsymbol{x}_t$ might have diverse units of measurement (thousands of people, billions of dollars, duration in weeks), resulting in large differences between their variances, which do not necessarily reflect differences in the information (that is, the reciprocal of the variance) they convey for forecasting y_{t+h}.

2.1. Classical Bayesian shrinkage

A noninformative prior, like the one assigned to the coefficients $\boldsymbol{\alpha}$, leads to a Bayes estimator centered at the unrestricted least squares (LS) quantities. This choice obviously poses a problem for estimating the "large" number of coefficients $\boldsymbol{\beta}$, especially when $p > T$. Traditionally, normal priors of the form

$$\pi(\boldsymbol{\beta}) \sim N(0, V) \tag{2}$$

have been used, because they are conjugate to the likelihood and allow easy calculations of the Bayes posterior. The $p \times p$ matrix V is the prior covariance matrix of the regression coefficients that we want to elicit for this "large p" problem. For instance, a common choice is the case $V = \tau^2 I_p$, which leads to the classical *ridge regression* shrinkage. Temporarily ignoring the effect of the regressors $\boldsymbol{z}_t$, this ridge regression prior implies the penalized least squares representation

$$\overline{\boldsymbol{\beta}}_{\text{ridge}} = \left(X'X + \frac{1}{\tau^2} I_p\right)^{-1} X'\boldsymbol{y},$$

where $X = \left(\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_T\right)'$ and $\boldsymbol{y} = (y_{1+h}, \ldots, y_{T+h})'$. The dependence of all parameters $\beta_j, j = 1, \ldots, p$, on the unknown parameter τ^2 can reduce the risk over the traditional LS estimator. For $\tau \to \infty$, we can see that $\boldsymbol{\beta} = \left(X'X\right)^{-1} X'\boldsymbol{y} = \boldsymbol{\beta}_{\text{LS}}$.

Following a different path, Judge and Bock (1978) suggested an empirical Bayes (i.e. data-based) estimator of V, of the form $V = \tau^2 \left(X'X\right)^{-1}$, where $\tau^2 = \frac{\widehat{\sigma}^2}{\widehat{\xi}^2}$, and

$$\widehat{\sigma}^2 = \left(\boldsymbol{y} - X\boldsymbol{\beta}'_{\text{LS}}\right)' \left(\boldsymbol{y} - X\boldsymbol{\beta}'_{\text{LS}}\right) / T$$

$$\widehat{\xi}^2 = \frac{\boldsymbol{\beta}'_{\text{LS}}\boldsymbol{\beta}_{\text{LS}}}{\operatorname{tr}\left(X'X\right)^{-1}} - \widehat{\sigma}^2.$$

This empirical Bayes rule is Stein-like, shrinking $\boldsymbol{\beta}_{\text{LS}}$ towards 0. This can be seen from the posterior mean, which can be written as a proportion of the LS estimator

$$\overline{\boldsymbol{\beta}}_{\text{EB}} = \left(1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2 + \widehat{\xi}^2}\right) \boldsymbol{\beta}_{\text{LS}}.$$

The more inaccurate the LS estimator is, the higher the model variance $\widehat{\sigma}^2$ is, meaning that the shrinkage factor $1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2 + \widehat{\xi}^2}$ tends to zero (leading the empirical Bayes posterior mean to approach zero as well).

Priors of the form $V = \sigma^2 \tau^2 \left(X'X\right)^{-1}$, which are called g-priors or Zellner priors (Zellner, 1986), tend to be very popular in economics; see for instance Koop and Potter (2004) and references therein. Over the years, there have been many connections made between values of τ^2 and information criteria; see for example the work of Fernandez, Ley, and Steel (2001) for a review. Nevertheless, Zellner originally proposed this prior for providing a shrinkage representation, since the posterior mean is of the form

$$\overline{\boldsymbol{\beta}}_{\text{g}} = \frac{\tau^2}{1 + \tau^2} \boldsymbol{\beta}_{\text{LS}}.$$

This formulation implies a shrinkage factor $\delta = \tau^2/(1 + \tau^2)$ that regulates the proportion (0%–100%) of shrinkage over the unrestricted OLS estimator.

We can immediately observe that these three typical examples of Bayesian shrinkage have undesirable properties for very demanding problems with many predictors. The ridge regression prior is based on a global shrinkage parameter τ^2 for all p regressors. In sparse regression problems, i.e. when p is very large and we expect that only a tiny proportion of regressors will be relevant for prediction, weighting all regression coefficients a priori by the same factor τ^2 is guaranteed not to work well. Empirical Bayes priors and g-priors partly solve this problem, since τ^2 is scaled by the Information Matrix, giving a varying degree of prior weight to each regression coefficient, based on the information in the likelihood. Nevertheless, the Information Matrix cannot be estimated precisely for large values of p, or be estimated at all for $p > T$.[3] Thus, the next subsection discusses shrinkage representations which are automatic (i.e., they allow minimal input by the researcher about the expected shrinkage factor), and can be applied in sparse regressions within the "large p, small T" paradigm.

2.2. Full Bayes (hierarchical) priors for adaptive shrinkage

Modern computational methods allow the prior covariance matrix V to be estimated in a formal way, by placing hyper-prior distributions on its elements. Using the Bayes theorem, the hyper-prior times the likelihood will provide an appropriate posterior expression for V, which in turn can be used as the prior for the coefficients of interest $\boldsymbol{\beta}$. Moreover, estimating V implies that we do not have to rely on using a single shrinkage factor τ for convenience, as is the case for the classical shrinkage priors discussed in the previous subsection. For instance, a popular expression for the prior covariance matrix of the coefficients $\boldsymbol{\beta}$ is $V = \operatorname{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$, which allows different coefficients β_j to shrink (when $\tau_j^2 \to 0$) through their prior to zero at a varying rate, while allowing others to remain unrestricted a priori (when $\tau_j^2 \gg 0$), for $j = 1, \ldots, p$. Even in the case of large values of p, having each element of V shrinking at a different rate is feasible, since each τ_j^2 will be updated "automatically" by the information in the data likelihood (through the Bayes theorem).

As a generic example, all of the cases examined in this paper can be cast into the prior form

$$\pi\left(\boldsymbol{\beta}|\tau^2\right) \sim N_p(0, V)$$

$$\pi\left(\tau_j^2\right) \sim F(\gamma),$$

[3] It is only recently that Maruyama and George (2010) derived a particular decomposition of Zellner's g-prior that can be used when more predictors than observations are present.

where $F(\gamma)$ denotes a generic prior distribution on τ_j^2 with hyperparameter(s) γ. In what follows, I analyze five popular specifications of $F(\gamma)$ which are commonly used in the literature, leading to five Bayesian shrinkage estimators, and connections with frequentist shrinkage estimators are highlighted. MCMC estimation and elicitation of the hyperparameter(s) is also discussed.

Adaptive shrinkage Jeffreys prior. Hobert and Casella (1996) studied the shrinkage properties of the Jeffreys prior on the covariance matrix of the regression coefficients. One can think of the Jeffreys prior as the simplest, default choice because it does not depend upon further hyperparameters. Let $V = \text{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$; then the scale-invariant, improper Jeffreys hyperprior on each τ_j^2 takes the form

$$\pi\left(\tau_j^2\right) \sim 1/\tau_j^2, \quad \text{for } j = 1, \ldots, p. \tag{3}$$

Adaptive shrinkage t-priors. For a covariance matrix $V = \text{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$, we can also consider a specific form of a gamma prior on $\tau_j^2, j = 1, \ldots, p$, i.e. the inverse gamma prior. Following Geweke (1993), we can show that this normal–inverse gamma mixture prior is equivalent to a Student-t prior on $\boldsymbol{\beta}$. The t-density has heavy tails and is more leptokurtic around the origin, which means that shrinkage around zero is achieved at a faster rate than for the simple normal prior. The priors on τ_j^2 are of the form

$$\pi\left(\tau_j^2\right) \sim \text{igamma}\left(\rho, \xi\right), \quad \text{for } j = 1, \ldots, p, \tag{4}$$

where ρ is the *shape* parameter and ξ the *scale* parameter of the inverse gamma density; see also Armagan and Zaretzki (2010). Once the τ_j^2s have been integrated out from the joint posterior, this prior is analogous to the regularized least squares problem which solves (once again ignoring the regressors $\boldsymbol{z}_t$, for simplicity)

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_{t+h} - \boldsymbol{x}_t'\boldsymbol{\beta}\right)^2 + \left(\rho + \frac{1}{2}\right)\sum_{j=1}^{p}\log\left(2\xi + \beta_j\right).$$

Finally, notice that this formula also applies to the Jeffreys prior case (which obtains when $\rho, \xi \to 0$).

Hierarchical lasso. Tibshirani (1996) proposed the lasso algorithm, which can be viewed as a L_1-penalized least squares estimate which solves

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_{t+h} - \boldsymbol{x}_t'\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|.$$

Tibshirani (1996) also noted that this form of penalized estimator is equivalent to the posterior mode of the Bayes estimate under the Laplace prior

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \sim \prod_{j=1}^{p}\frac{\lambda}{2\sqrt{\sigma^2}}e^{-\frac{\lambda}{\sqrt{\sigma^2}}\left|\beta_j\right|}.$$

One can take advantage of the fact that the Laplace density can be written as a scaled mixture of normals (see Park & Casella, 2008). Note that the formulation above implies that, for the Bayesian lasso prior (as well as the fused lasso and the elastic net discussed below), we need to condition on the error variance σ^2. Park and Casella (2008) underline the fact that this conditioning ensures that the posterior of the regression coefficients $\boldsymbol{\beta}$ is unimodal, otherwise expensive simulation methods (for instance, simulated tempering) would be needed to handle multimodal posteriors. Subsequently, assume for this case a diagonal prior covariance matrix of the form $V = \sigma^2 \times \text{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$. The hierarchical version of the lasso uses a normal prior for $\boldsymbol{\beta}$ of the form in Eq. (2), augmented with the hyperprior

$$\pi\left(\tau_j^2|\lambda\right) \sim \text{exponential}\left(\frac{\lambda^2}{2}\right), \quad \text{for } j = 1, \ldots, p, \tag{5}$$

where λ is a hyperparameter, which is the *rate* parameter of the exponential distribution.

Hierarchical fused lasso. The fused lasso was proposed by Tibshirani et al. (2005) as a means of accounting for any possible meaningful ordering of variables.[4] This estimator penalizes the L_1-norm of both the coefficients and their differences

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_{t+h} - \boldsymbol{x}_t'\boldsymbol{\beta}\right)^2 + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p-1}\left|\beta_{j+1} - \beta_j\right|.$$

The representation of the Bayesian prior for $\boldsymbol{\beta}$ in the penalized regression using the fused lasso is

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \sim e^{-\frac{\lambda_1}{\sigma}\sum_{j=1}^{p}\left|\beta_j\right| - \frac{\lambda_2}{\sigma}\sum_{j=1}^{p-1}\left|\beta_{j+1}-\beta_j\right|}.$$

Kyung et al. (2010) show that the hierarchical representation of this prior is

$$\pi\left(\tau_j^2|\lambda_1\right) \sim \text{exponential}\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, \ldots, p, \tag{6a}$$

$$\pi\left(\omega_j^2|\lambda_2\right) \sim \text{exponential}\left(\frac{\lambda_2^2}{2}\right), \quad \text{for } j = 1, \ldots, p-1, \tag{6b}$$

where the correlation between β_{j+1} and β_j enters through the prior covariance matrix V. In this case, V is a tridiagonal matrix, with main diagonal $\left\{\tau_i^2 + \omega_{i-1}^2 + \omega_i^2\right\}$ for $i = 1, \ldots, p$, and off-diagonal elements $\left\{-\omega_i^2\right\}$, and for simplicity we can set $\omega_0 = \omega_p = 0$.

Hierarchical elastic net. Zou and Hastie (2005) proposed the elastic net as a more stabilized version of the lasso that also allows grouping effects and is particularly useful when $p > T$. The elastic net estimator is the solution to the minimization problem

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_{t+h} - \boldsymbol{x}_t'\boldsymbol{\beta}\right)^2 + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2.$$

[4] The data set in this paper implies such an ordering. Many disaggregated and component series of the same aggregated series appear in order. In addition, all variables in this dataset are ordered according to statistical releases.

A Bayesian prior for $\boldsymbol{\beta}$ in the penalized regression using this estimator is

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \sim e^{-\frac{\lambda_1}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2}.$$

Kyung et al. (2010) show that a hierarchical representation of this density exists, and that it is of double-exponential form, as in the simple lasso. This means that the hyperprior on τ_j^2 is

$$\pi\left(\tau_j^2|\lambda_1^2\right) \sim \text{exponential}\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, \ldots, p, \quad (7a)$$

where, in this case, the difference from the standard lasso prior is that the covariance matrix is of the form $V = \sigma^2 \times \text{diag}\left\{\left(\tau_1^{-2} + \lambda_2\right)^{-1}, \ldots, \left(\tau_p^{-2} + \lambda_2\right)^{-1}\right\}$.

As opposed to maximizing the likelihood using no prior information, estimation for the Bayesian means that the likelihood function is multiplied by a well defined prior distribution (in our case, a shrinkage prior). The resulting posterior distribution of the regression coefficients, $\boldsymbol{\beta}$, conveys all we need to know about our model. For instance, the mode of the posterior is identical to the corresponding frequentist shrinkage estimators. In Appendix B, I give all of the details necessary for obtaining samples from the posterior distribution of all regression parameters $\boldsymbol{\theta} = \left\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2\right\}$ by sampling from their conditional posteriors using Markov Chain Monte Carlo (MCMC) methods.

In particular, the Gibbs sampler is used to obtain draws from the posterior $p(\theta)$ by sampling from the conditional densities $p\left(\boldsymbol{\alpha}|\boldsymbol{\beta}, \sigma^2\right)$, $p\left(\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2\right)$ and $p\left(\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}\right)$. Even when $p(\boldsymbol{\theta})$ is a distribution that is difficult to sample from, it turns out that the posterior conditionals are well-known distributions that are easy to sample from (normal, gamma, inverse Gaussian, and so on). Hobert and Geyer (1998) proved the geometric ergodicity of the two-stage Gibbs sampler from hierarchical models of a general normal-gamma form, a result that can be generalized to the lasso, fused lasso and elastic net priors (see Kyung et al., 2010).

A general MCMC algorithm accommodating all five of the shrinkage priors described above involves sampling iteratively from the following conditional posteriors of the regression parameters $\boldsymbol{\theta} = \left\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2\right\}$, plus the hyperparameters τ_j^2 and λ:

1. Sample $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^{p}, \lambda$.
2. Sample τ_j^2 for $j = 1, \ldots, p$ from $\tau_j^{-2}|\boldsymbol{\beta}, \sigma^2, \lambda$.
3. Sample λ from $\lambda^2|\left\{\tau_j^2\right\}_{j=1}^{p}$.
4. Sample $\boldsymbol{\alpha}$ from $\boldsymbol{\alpha}|\boldsymbol{\beta}, \sigma^2$.
5. Sample σ^2 from $\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^{p}$.
6. Repeat Steps 1–5 using the most recent values of the conditioning variables, where we skip step 3 (this hyperparameter does not exist) for the Jeffreys and Student-t priors, while $\lambda = (\lambda_1, \lambda_2)$ in the fused lasso and elastic net algorithm.

Note that $\boldsymbol{\alpha}$ need not be sampled conditional on $\boldsymbol{\beta}$, and vice-versa; instead, these vectors of regression coefficients could be sampled together from a $(p+q)$-variate normal density, using data $\boldsymbol{w}_t = (\boldsymbol{z}_t, \boldsymbol{x}_t)$. In this application, however, since $\boldsymbol{\alpha}$ contains only the intercept and coefficients on two lags, sampling $\boldsymbol{\alpha}|\boldsymbol{\beta}$ and $\boldsymbol{\beta}|\boldsymbol{\alpha}$ should not decrease the sampling efficiency too significantly, relative to the joint sampling of both parameter vectors.

2.2.1. *Tuning the hyperparameters*

Hierarchical priors have the advantage of allowing the data to determine the prior hyperparameter of interest (covariance matrix of the normal prior in our case). However, from the formulations above, we can observe that introducing a second layer of hierarchy (the gamma-type densities) means that at least one new hyperparameter is introduced; it is only for the normal-uniform prior that this is obviously not the case. For the normal–inverse gamma prior (Student-t), we need to select values for the hyperparameters (ρ, ξ) of the inverse gamma density. Although one can easily set a prior on the scale parameter ξ,[5] typical noninformative values for the inverse gamma distribution in Bayesian analysis are usually $\rho = \xi = 0.01$ or $\rho = \xi = 0.001$ (see Gelman, 2006). Since the inverse gamma becomes equivalent to a Jeffreys prior for τ_j^2 (which is the first shrinkage prior examined) for these very low values of (ρ, ξ), I will examine the more informative prior igamma ($\rho = 3, \xi = 0.001$), which concentrates τ_j^2 around the neighborhood of zero (note that the variance of the inverse gamma does not exist for $\rho \leq 2$).

For the hierarchical lasso prior, the regularization parameter λ controls the intensity of the penalty in each regression coefficient, and it should ideally be updated by the data likelihood. A conjugate prior which would facilitate posterior computations is the gamma prior on λ^2 (not λ), of the form

$$\pi\left(\lambda^2\right) \sim \text{gamma}\left(r, \delta\right).$$

Similarly, an additional layer on the hyperparameters λ_1, λ_2 of the fused lasso and elastic net priors is of the form

$$\pi\left(\lambda_1^2\right) \sim \text{gamma}\left(r_1, \delta_1\right)$$

$$\pi\left(\lambda_2^2\right) \sim \text{gamma}\left(r_2, \delta_2\right),$$

and, hence, it is easy to verify that setting $r = \delta = 0.01$ (and similarly $r_1 = \delta_1 = r_2 = \delta_2 = 0.01$) will produce near-uniform (noninformative) priors on the hyperparameters λ, λ_1 and λ_2.

3. Empirical results

The data-set consists of 129 quarterly U.S. macroeconomic time series spanning the period 1959:Q1 to 2010:Q2 (the effective sample size, after converting to stationarity and specifying lags, is 1960:Q1–2010:Q2). The series were downloaded from the St. Louis Fed FRED database,[6] and a complete description is given in Table A.1 of the Appendix A. The whole data set is quite standard for this type of application, and includes data releases such as personal income and outlays, GDP and components, assets and liabilities of commercial banks in the United States, productivity and costs measures, exchange rates and selected interest rates, among others. All series are seasonally adjusted, where this is applicable, and transformed to be

[5] A conjugate prior on ξ is the Gamma $\left(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0\right)$ density. Then, the posterior of the inverse Gamma prior is again an inverse Gamma density with parameters $\left(\boldsymbol{\alpha}_0 + p\rho, \boldsymbol{\beta}_0 + \sum_{i=1}^{p}\tau_i^{-2}\right)$.

[6] http://research.stlouisfed.org/fred2/.

stationary. When the series are used as predictors in $\boldsymbol{x}_t$, standard stationarity transformations are applied, like first and second (log) differences. In contrast, when the series are used as the series to be predicted (y_{t+h}), then h-quarter growth or differences transformations are used. All transformations are summarized in column "T" in Table A.1 and explained in detail in Appendix A (see also Table A.2).

As was mentioned in the Introduction, the factor model is used as a benchmark shrinkage estimator. Details of the exact definitions of the factors are given in the following subsection. An important consideration when extracting factors is that there are series in the data set which are higher level aggregates (mainly sums) of individual disaggregated series. As it is not sensible to extract a common factor between, say, two series and their sum, 14 such aggregate series in the data set are excluded from the factor estimation. In column "F" in Table A.1, the 115 disaggregated variables which are used to extract factors are denoted by 1. This restriction does not hold when the shrinkage priors and all series are used as predictors.

3.1. *Forecasting with many predictors*

All forecasts are from the univariate regression in Eq. (1), where I use one of the 129 variables (iteratively) as the dependent variable (y_{t+h}), and the remaining 128 variables enter the regression as the matrix of standardized exogenous predictors $(\boldsymbol{x}_t)$, which are subject to shrinkage. Then, the five Bayesian shrinkage priors are applied to estimate $\widehat{\boldsymbol{\beta}}_j$ (j = *Jeffreys, Student-t, Lasso, Fused Lasso, Elastic Net*), and forecasts are produced using the original, unstandardized matrix of predictors $\boldsymbol{x}_t^*$. In order to forecast with the factor regression model, $\boldsymbol{x}_t$ is replaced with the first five principal components of the 115 disaggregated series in $\boldsymbol{x}_t$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{DFM}}$ is estimated by simple OLS. The variables which are always included in each of the six forecasting models (i.e., variables in $\boldsymbol{z}_t$ which are unrestricted) are the intercept and two lags of the one-quarter growth rates or differences of the dependent variable (i.e., lags y_t and y_{t-1} using the same stationarity transformations as in the variables in $\boldsymbol{x}_t$).

The first estimation period (after taking lags and transforming to stationarity) is 1960:Q1 to 1984:Q4, and the sample 1985:Q1 to 2010:Q2 (last 102 observations) is kept for the evaluation of h-step-ahead forecasts, $h = 1, 2, 4$. Specifically, using the initial sample (where y_{t+h} is observed from 1960:Q1 + h to 1984:Q4 and $(\boldsymbol{z}_t, \boldsymbol{x}_t)$ is observed from 1960:Q1 to 1984:Q4 − h), estimation of the regression in Eq. (1) provides parameter estimates $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2$, and then forecasts can be computed for $y_{1984:Q4+h}$ by plugging into the regression the realization of the predictors in 1984:Q4, i.e., the values $\left(\boldsymbol{z}_{1984:Q4}, \boldsymbol{x}_{1984:Q4}\right)$. Then one data point is added and the same procedure is followed, until the sample is exhausted. Since the models with shrinkage priors are estimated using Monte Carlo, which provides draws from the whole posterior density of the parameters, we can use the same method to obtain the full predictive density of the regression model. For each of the 129 dependent variables, five prior distributions, three forecast horizons, and $102-h$ out-of-sample observations, we save 7000 post-burn-in draws from the conditional posteriors of the regression parameters $\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2\right)$ (see Appendix B for the exact formulæ), and generate 10 forecasts using each parameter draw,[7] leading to 70,000 draws from the predictive density of each of the 129 variables.

In order to be able to judge how well the chain mixes with just 7000 post-burn-in draws, Fig. 1 provides boxplots of the inefficiency factors (i.e., inverse of the relative numerical efficiency measure of Geweke, 1993) associated with specific parameters. The inefficiency factor is an estimate of $1 + 2\sum_{k=1}^{\infty} \rho_k$, with ρ_k denoting the k-th order autocorrelation of the draws. The estimate is performed using a 4% tapered window for the estimation of the spectral density at frequency zero. As a rule of thumb, values of the inefficiency factors below 20 are regarded as satisfactory. In Fig. 1, the boxplots are constructed using the inefficiency factors of all of the regression coefficients, $\boldsymbol{\beta}$, and the regularization parameter(s), λ, in the 129 estimated regression models, and for the case of the lasso and elastic net priors.[8] For some regressions, the inefficiency factors on λ_1 and λ_2 of the elastic net estimator are quite high. However, the estimation of the same regressions using 50,000 iterations does not alter the value of the mean forecasts, which is the main focus of this paper.

In a similar comparison of shrinkage estimators for regressions with many predictors, Stock and Watson (2011) use 4 lags in each of their 143 univariate regressions, and report their results relative to an AR(4) model. De Mol et al. (2008) consider only an unrestricted intercept in their shrinkage regressions, and report their results relative to a random walk. In this paper, since the effects of an intercept and two lags are removed in each forecasting model using uninformative priors, it is natural to consider forecast performance statistics relative to an AR(2) model estimated with unrestricted least squares. In this paper, Mean Absolute Forecast Errors (MAFE) and Mean Squared Forecast Errors (MSFE) are considered. Unless otherwise stated, all results are based on the MAFE and MSFE statistics of model j relative to the MAFE and MSFE of the AR(2) model (i.e., $\mathrm{MAFE}_j = \mathrm{mafe}_j/\mathrm{mafe}_{AR(2)}$ and $\mathrm{MSFE}_j = \mathrm{msfe}_j/\mathrm{msfe}_{AR(2)}$). Consequently, $\mathrm{MAFE}_j > 1$ means that the AR(2) dominates in terms of absolute forecast error, while the opposite is true for $\mathrm{MAFE}_j < 1$.

Tables 1–3 present the average absolute forecast errors for 1, 2 and 4 quarters ahead. Since the relative MAFE results are averaged over many series, three decimal places are used in this table because the differences are quite small otherwise (see also Stock & Watson, 2011). First, based on the median MAFE using the total number of series, the simple lasso and the elastic net give the smallest

[7] Since it is costly to use more iterations in order to reduce the sampling error in the posterior of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and σ^2 ("parameter sampling error"), it is computationally easier to decrease the sampling error associated with the specific model used ("model sampling error"), which in effect results in a reduction of the sampling error of the predictive density.

[8] Note that $\boldsymbol{\beta}$ has $p = 128$ elements in each of the 129 regressions, but, for space reasons, one boxplot is constructed using all 128 × 129 inefficiency factors on $\boldsymbol{\beta}$ (the individual boxplots for each β_i are not that different).

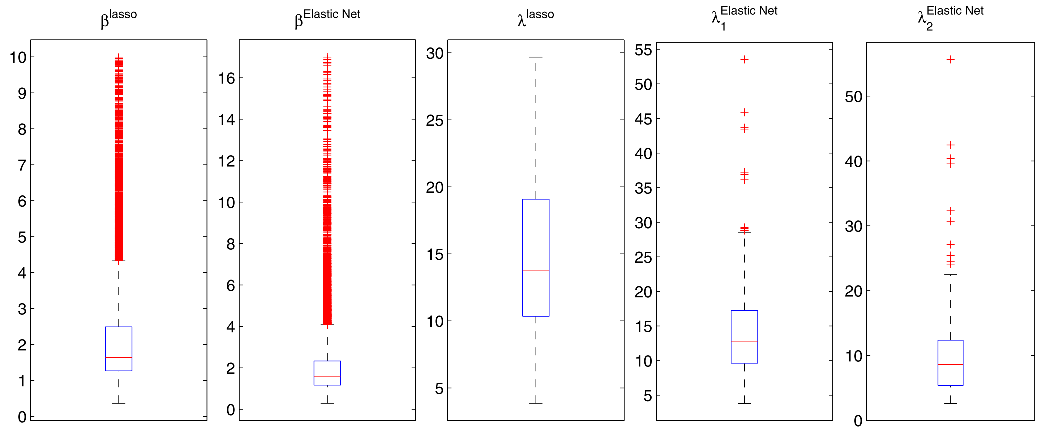

Fig. 1. Boxplots of inefficiency factors for the regression coefficients, $\boldsymbol{\beta}$, and the regularization parameter(s), λ, for the lasso and elastic net priors when using 7000 draws. These inefficiency factors are estimated across 129 predictive regressions, with a forecast horizon of $h = 1$.

Table 1
MAFE results for the five Bayes shrinkage estimators and the DFM, $h = 1$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net	DFM
Median MAFEs based on statistical releases						
GDP and components	1.064	1.001	0.987	1.016	**0.986**	1.030
Housing	1.255	1.020	1.039	1.019	1.040	**0.924**
IP	1.016	1.027	0.988	1.013	0.988	**0.928**
Employment situation	1.058	1.024	1.012	1.033	1.012	**0.976**
Productivity/Costs	1.018	1.169	0.986	1.043	0.987	**0.959**
Assets/Liabilities of banks	0.966	0.973	0.970	0.973	0.971	**0.957**
Interest rates	1.056	0.970	0.967	0.991	**0.966**	1.066
Money stock	0.917	0.910	0.905	0.908	**0.904**	1.050
Currency in circulation	0.677	0.680	0.678	**0.671**	0.677	0.678
MZM	0.910	0.894	0.896	0.900	**0.893**	1.294
Money velocity	0.994	0.976	0.981	0.990	**0.979**	1.092
Consumer credit	1.001	0.985	0.985	0.998	**0.980**	1.043
CPI	1.006	0.914	0.916	0.914	**0.909**	0.963
PPI	**0.913**	0.917	0.916	0.918	0.919	0.931
Stock prices	**1.003**	1.008	1.006	1.004	1.005	1.072
Exchange rates	0.999	1.004	1.003	**0.997**	1.003	0.999
ISM surveys	1.108	1.276	1.031	1.038	1.028	**0.987**
MAFE descriptives based on total number of series						
median	1.017	0.998	**0.987**	1.004	**0.987**	0.993
variance	0.045	0.509	0.008	0.015	0.007	0.015
min	0.677	0.680	0.678	0.671	0.677	0.678
max	1.857	2.342	1.435	1.645	1.286	1.582

Note: Entries are MAFE-based statistics, relative to the MAFE of an AR(2) model.

forecast errors in all cases (note that there is a difference for $h = 2$, but it is minimal). This might suggest that taking the correlation among the predictors into account, which is what the elastic net algorithm adds to the simple lasso algorithm, is not that important with these data. However, the elastic net consistently has the smallest maximum MAFE, and consequently has the smallest variance across all 129 series. The same shrinkage algorithm clearly dominates in most of the 17 data categories for $h = 1$, on average. For $h = 4$, the three shrinkage estimators (lasso, elastic net and the DFM) all do equally well.

Hierarchical shrinkage priors based on the uniform and inverse gamma distributions perform very poorly on average, although they provide the smallest MAFE values among all shrinkage estimators for some data categories. Looking at the MAFE descriptives based on the total number of series, Student-*t* shrinkage always performs better than Jeffreys shrinkage in lowering the median MAFE. However, note that for the Student-*t* prior, a single default choice of hyperparameters is applied to all 129 series. Although this choice works well on average (median MAFEs), it performs poorly in some series. For example, there are cases where this estimate leads to MAFEs which are up to ten times higher (see the maximum MAFE based on the total number of series in Table 3) than those of the benchmark model. On the other hand, the Jeffreys prior does not depend on the choice of hyperparameters, and we can safely say that its shrinkage and forecasting performance is very unsatisfactory for the specific design of this study. Finally, the idea behind the fused lasso, i.e., taking into account the correlation among consecutive predictors, does not help improve the forecasting performance at all. In fact, forecasts from this estimator are always dominated by the lasso and the elastic net.

Table 2
MAFE results for the five Bayes shrinkage estimators and the DFM, $h = 2$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net	DFM
Median MAFEs based on statistical releases						
GDP and components	1.100	1.025	**0.988**	1.025	**0.988**	1.005
Housing	1.208	1.045	1.040	1.035	1.043	**1.017**
IP	1.050	1.086	**0.987**	1.026	0.989	0.989
Employment situation	1.089	1.029	1.017	1.039	1.016	**0.982**
Productivity/Costs	1.160	1.386	1.081	1.184	1.082	**0.942**
Assets/Liabilities of banks	**0.967**	1.072	0.973	0.977	0.969	0.968
Interest rates	1.294	0.998	**0.986**	1.082	0.987	1.000
Money stock	**0.929**	0.933	0.932	0.930	0.931	1.002
Currency in circulation	0.737	0.736	0.737	**0.733**	0.735	0.746
MZM	0.855	0.855	**0.845**	0.847	0.849	1.112
Money velocity	1.000	**0.995**	0.996	1.002	0.998	1.029
Consumer credit	**0.990**	0.999	0.996	**0.990**	0.996	1.010
CPI	1.080	0.962	0.955	0.979	**0.954**	0.994
PPI	0.958	0.942	**0.941**	0.957	0.945	0.975
Stock prices	**0.987**	0.995	0.997	0.992	0.996	1.049
Exchange rates	1.026	1.007	1.008	**1.005**	1.009	1.030
ISM surveys	1.041	1.415	**1.021**	1.057	**1.021**	1.050
MAFE descriptives based on total number of series						
median	1.038	1.015	**0.990**	1.019	0.991	0.999
variance	0.296	0.625	0.009	0.029	0.008	0.007
min	0.491	0.494	0.487	0.495	0.482	0.488
max	3.230	4.536	1.458	1.931	1.289	1.222

Note: Entries are MAFE-based statistics, relative to the MAFE of an AR(2) model.

Table 3
MAFE results for the five Bayes shrinkage estimators and the DFM, $h = 4$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net	DFM
Median MAFEs based on statistical releases						
GDP and components	1.094	1.061	**0.990**	1.022	**0.990**	0.994
Housing	1.043	1.071	1.061	1.062	1.062	**0.979**
IP	1.030	1.040	**0.983**	1.022	0.987	0.997
Employment situation	1.188	1.029	1.004	1.075	1.004	**0.937**
Productivity/Costs	1.320	1.544	1.112	1.397	1.109	**0.891**
Assets/Liabilities of banks	**0.968**	1.128	0.979	0.998	0.970	0.969
Interest rates	1.624	1.050	1.002	1.113	1.000	**0.969**
Money stock	**0.933**	**0.933**	0.935	0.934	0.938	1.019
Currency in circulation	0.916	0.920	0.918	0.925	0.920	**0.884**
MZM	0.928	0.931	0.930	**0.922**	0.934	1.117
Money velocity	1.000	0.987	**0.985**	0.993	0.987	1.013
Consumer credit	0.967	1.006	**0.964**	0.990	0.968	1.015
CPI	1.190	1.698	**0.964**	1.051	0.968	1.031
PPI	0.983	0.964	0.964	0.989	**0.962**	1.010
Stock prices	0.968	**0.948**	0.952	0.951	0.959	1.031
Exchange rates	1.093	1.053	1.029	**1.027**	1.033	1.036
ISM surveys	1.045	1.289	**0.982**	1.026	0.984	0.990
MAFE descriptives based on total number of series						
median	1.046	1.029	**0.989**	1.025	**0.989**	**0.989**
variance	0.693	1.776	0.009	0.081	0.007	0.009
min	0.578	0.575	0.576	0.574	0.578	0.610
max	6.707	9.883	1.609	3.386	1.364	1.275

Note: Entries are MAFE-based statistics, relative to the MAFE of an AR(2) model.

Once we turn to Tables 4–6 with the MSFE results based on the total number of series, it is obvious that the DFM is dominating all Bayesian shrinkage estimators at all three forecast horizons. Although the lasso and the elastic net improve over the benchmark AR(2) forecasts, they are still not as good as the DFM. Nevertheless, by looking at the individual data categories, the elastic net is the best in forecasting GDP and its components at horizons $h = 1, 2$, as well as the various Consumer Price Indexes at all forecast horizons. Note that this pattern was also true for the MAFE results in Tables 1–3. Therefore, summarizing the results in Tables 1–6, the elastic net and the lasso are the best Bayesian shrinkage estimators, from a mean forecast error point of view. However, these might not produce too much improvement over principal component shrinkage using a factor model, and the final result is dependent on the series being forecasted each time.

Table 7 gives a better view of the total performance of each shrinkage estimator. Hit rates are calculated based on MAFEs, MSFEs and predictive likelihoods. These are estimated as the proportion of times (among the 129 series) that a specific shrinkage estimator had the lowest MAFE, the lowest MSFE and the highest average predictive likelihood (APL). The average predictive likelihood can be

Table 4
MSFE results for the five Bayes shrinkage estimators and the DFM, $h = 1$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net	DFM
Median MSFEs based on statistical releases						
GDP and components	1.036	0.996	0.993	1.002	**0.994**	1.047
Housing	1.236	1.068	1.059	1.059	1.059	**0.974**
IP	1.017	1.022	0.998	1.024	1.000	**0.843**
Employment situation	1.036	1.026	1.019	1.037	1.019	**0.899**
Productivity/Costs	1.003	1.108	0.986	1.023	0.985	**0.896**
Assets/Liabilities of banks	0.970	0.973	0.969	0.973	0.972	**0.960**
Interest rates	1.045	**0.954**	0.959	0.995	0.956	0.995
Money stock	0.948	0.949	0.945	0.946	**0.943**	1.094
Currency in circulation	0.746	0.747	0.750	0.748	0.747	**0.563**
MZM	0.920	0.905	0.907	**0.904**	**0.904**	1.668
Money velocity	1.009	0.994	0.996	1.007	**0.994**	1.129
Consumer credit	0.989	0.978	**0.976**	0.985	**0.976**	1.087
CPI	0.995	0.940	**0.936**	0.945	**0.936**	0.945
PPI	0.939	0.947	0.942	0.945	0.945	**0.906**
Stock prices	1.007	1.007	1.006	1.006	**1.003**	1.133
Exchange rates	0.999	1.002	0.999	1.000	**0.998**	1.016
ISM surveys	1.091	1.236	1.026	1.016	1.025	**0.932**
MSFE descriptives based on total number of series						
median	1.007	0.994	0.991	1.002	0.991	**0.958**
variance	0.041	0.084	0.006	0.010	0.006	0.054
min	0.671	0.747	0.666	0.706	0.664	0.452
max	2.380	2.797	1.474	1.545	1.370	2.498

Note: Entries are MSFE-based statistics, relative to the MSFE of an AR(2) model.

Table 5
MSFE results for the five Bayes shrinkage estimators and the DFM, $h = 2$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net	DFM
Median MSFEs based on statistical releases						
GDP and components	1.074	1.028	1.004	1.022	**0.998**	1.004
Housing	1.168	1.082	1.072	1.074	**1.071**	1.161
IP	1.041	1.152	1.012	1.049	**1.011**	1.013
Employment situation	1.055	1.035	1.020	1.036	1.022	**0.985**
Productivity/Costs	1.128	1.327	1.075	1.155	1.077	**0.882**
Assets/Liabilities of banks	0.976	1.037	0.980	0.970	0.977	**0.912**
Interest rates	1.260	1.008	0.994	1.084	0.998	**0.946**
Money stock	0.974	0.974	0.976	**0.972**	0.973	1.062
Currency in circulation	0.785	0.783	0.786	0.788	0.785	**0.627**
MZM	0.915	0.911	0.907	**0.904**	0.910	1.392
Money velocity	1.021	1.003	**1.002**	1.019	1.005	1.026
Consumer credit	0.986	0.994	**0.991**	0.996	0.993	0.993
CPI	1.048	0.967	0.966	0.986	**0.965**	0.999
PPI	0.979	0.968	**0.967**	0.976	0.968	0.975
Stock prices	**0.994**	1.000	1.000	1.000	1.000	1.073
Exchange rates	1.020	1.012	1.011	**1.010**	1.011	1.023
ISM surveys	1.036	1.697	**1.007**	1.020	1.010	1.043
MSFE descriptives based on total number of series						
median	1.035	1.018	0.998	1.017	0.999	**0.995**
variance	0.144	0.205	0.007	0.017	0.006	0.026
min	0.563	0.568	0.557	0.562	0.554	0.304
max	3.392	3.536	1.420	1.667	1.313	1.897

Note: Entries are MSFE-based statistics, relative to the MSFE of an AR(2) model.

used to evaluate the whole predictive density of each regression model; see Geweke and Amisano (2010) and references therein. The predictive likelihood (or marginal posterior likelihood) is simply the predictive density of the regression model evaluated at the out-of-sample observation y_{t+h}. Although Tables 1–6 showed the elastic net having exactly the same median MAFE and MSFE as the lasso based on the total number of series, Table 7 shows that the lasso has better hit rates for all three measures. In terms of mean forecasting, the use of the lasso leads to even better improvements in the density forecasts than the use of the elastic net. In terms of density forecasting, the lasso improves the density forecasts from the elastic net even more (an average improvement of 25% at all forecast horizons). This is because parameter uncertainty feeds into the predictive likelihood evaluation. Thus, the elastic net having two regularization parameters λ_1 and λ_2, the uncertainty (posterior variance) about the two parameters feeds into the density forecasts of y_{t+h}. The lasso, having only one regularization parameter, i.e., $\lambda_1 = \lambda$ and $\lambda_2 = 0$, has a lower forecast uncertainty/variance (given that the forecasts of the mean coming from the two estimators are more or less identical for this specific case-study).

Table 6
MSFE results for the five Bayes shrinkage estimators and the DFM, $h = 4$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net	DFM
Median MSFEs based on statistical releases						
GDP and components	1.061	1.057	**0.997**	1.022	0.999	1.004
Housing	**1.065**	1.097	1.087	1.092	1.087	1.106
IP	1.016	1.025	0.996	1.023	0.998	**0.986**
Employment situation	1.112	1.025	1.006	1.048	1.005	**0.916**
Productivity/Costs	1.222	1.614	1.099	1.289	1.093	**0.815**
Assets/Liabilities of banks	0.975	1.094	0.977	0.988	0.975	**0.924**
Interest rates	1.571	1.084	1.001	1.096	1.003	**0.957**
Money stock	**0.973**	0.974	0.975	**0.973**	0.975	1.114
Currency in circulation	0.905	0.906	0.905	0.908	0.907	**0.782**
MZM	0.932	0.934	0.934	**0.927**	0.938	1.194
Money velocity	1.020	0.999	0.996	1.013	0.998	**0.987**
Consumer credit	**0.969**	1.026	0.971	0.987	0.973	1.032
CPI	1.085	1.354	0.969	1.014	**0.968**	1.081
PPI	0.972	**0.962**	0.964	0.977	0.963	0.999
Stock prices	0.984	0.976	0.978	0.979	0.981	1.090
Exchange rates	1.098	1.060	**1.035**	**1.035**	1.036	1.062
ISM surveys	1.052	1.521	0.998	1.032	**0.996**	1.026
MSFE descriptives based on total number of series						
median	1.039	1.022	0.995	1.026	0.996	**0.987**
variance	0.386	0.417	0.008	0.047	0.006	0.028
min	0.620	0.610	0.611	0.612	0.612	0.414
max	5.389	6.784	1.573	2.771	1.378	1.620

Note: Entries are MSFE-based statistics, relative to the MSFE of an AR(2) model.

Table 7
Hit-rates of the five Bayes estimators, total number of series.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net
Hit rates, $h = 1$					
% of lowest MAFE	14.0	17.1	**35.7**	10.9	22.5
% of lowest MSFE	15.5	20.9	**34.9**	7.8	20.9
% of highest APL	0.8	2.3	**62.0**	7.8	27.1
Hit rates, $h = 2$					
% of lowest MAFE	18.6	13.2	**34.1**	11.6	22.5
% of lowest MSFE	17.1	20.9	**28.7**	11.6	21.7
% of highest APL	1.6	0.8	**60.5**	13.9	23.3
Hit rates, $h = 4$					
% of lowest MAFE	17.8	12.4	**31.8**	10.9	27.1
% of lowest MSFE	13.2	17.1	**34.9**	10.1	24.8
% of highest APL	0.0	0.0	**55.8**	8.5	35.7

Note: This table shows the proportion of times (over the 129 series being forecasted) that each estimator achieved the lowest value of the MAFE and MSFE statistics, and the highest value of the Average Predictive Likelihood (APL).

Table 8
Average similarity of Bayes forecasts, $h = 1$: correlation (lower left) and mean absolute difference of forecasts (upper right).

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net
Jeffreys		0.088	0.037	0.032	0.037
Student-*t*	0.944		0.080	0.083	0.080
Lasso	1.000	0.944		0.015	0.002
Fused lasso	1.000	0.945	1.000		0.015
Elastic net	1.000	0.944	1.000	1.000	

The similarity among the five shrinkage forecasts is assessed in Table 8. The lower triangular entries in this table show the correlation coefficients of all MSFEs for all 129 variables for horizon $h = 1$. The correlations among all shrinkage forecast errors are all one, except for the Student-*t* forecasts, which are less correlated with the other four shrinkage methods. This is simply because the other four hierarchical shrinkage priors (Jeffreys, lasso, fused lasso and elastic net) are based on noninformative priors on the lower level of their hierarchies. The entries above the diagonal of Table 8 are the mean absolute differences in the mean forecast errors between the prior mentioned in the row of the table and the prior mentioned in its column, averaged across all 129 series. The results confirm that the Student-*t* forecasts, which were the worst according to Tables 1–6, are the most distant from the forecasts generated from the other four priors. In contrast, the lasso and elastic net forecasts have the smallest

Table 9
MAFEs, MSFEs and Predictive Likelihoods for all 129 series, orthogonal predictors, $h = 1$.

	Jeffreys	Student-*t*	Lasso	Fused lasso	Elastic net
MAFE descriptives based on the total number of series					
median	0.9876	0.9871	0.9873	0.9846	0.9863
25% quantile	0.9539	0.9536	0.9523	0.9551	0.9505
75% quantile	1.0166	1.0128	1.0183	1.0139	1.0167
variance	0.0061	0.0068	0.0060	0.0062	0.0061
min	0.6768	0.6812	0.6813	0.6800	0.6821
max	1.2988	1.3151	1.2797	1.3053	1.2803
MSFE descriptives based on the total number of series					
median	0.9901	0.9923	0.9901	0.9903	0.9904
25% quantile	0.9558	0.9558	0.9559	0.9548	0.9510
75% quantile	1.0156	1.0227	1.0182	1.0178	1.0175
variance	0.0046	0.0049	0.0046	0.0048	0.0047
min	0.6714	0.6726	0.6666	0.6676	0.6670
max	1.2197	1.2750	1.2150	1.2397	1.2160
PL descriptives based on the total number of series					
median	0.4190	0.4175	0.4711	0.4837	0.4724
25% quantile	0.2241	0.2232	0.2636	0.2717	0.2636
75% quantile	0.7112	0.7050	0.7839	0.7968	0.7854
variance	0.2065	0.1679	0.1867	0.1954	0.1876
min	0.0380	0.0379	0.0487	0.0510	0.0487
max	3.7777	3.2578	3.3826	3.4491	3.3998

differences among all other pairs of methods, something which was also confirmed by their equal forecasting performance, as shown in Tables 1–6.

No matter how correlated the forecasts from the different shrinkage priors are on average, we saw in Tables 1–6 that their differences across different data releases and across forecasts horizons were substantial. The lasso and elastic net priors have a better ability to take into account the correlation patterns in the predictor variables, while the fused lasso is less good at this task (because of the very specific correlation pattern it has to find, i.e., penalize fewer/more consecutive predictors as a group). The Jeffreys and Student-*t* priors do not account for correlation in the predictor variables explicitly, and hence, their performance can be very risky at times, with forecast errors that are multiples of those produced by the other three methods.

If the correlation among the predictors is a crucial determinant of the performance of these algorithms, then it is natural to ask what happens if we forecast with orthogonal predictors (the case that is examined by Stock & Watson, 2011). Table 9 presents MAFE, MSFE and APL descriptive statistics based on all 129 series for $h = 1$, when the exogenous predictors are orthogonalized. For that purpose, the MATLAB function ORTH is used, which creates an orthonormal basis for the range of the matrix of exogenous predictors $\boldsymbol{x}_t$, and which is based on simply taking the singular value decomposition of $\boldsymbol{x}_t$. In fact, as is seen in Table 9, the orthogonalization of the data amounts to an almost identical forecasting performance of the five Bayesian algorithms. In addition, their performance is equal to that of the best performing method, the lasso, when using correlated predictors (compare the total MAFE and total MSFE results in Tables 1 and 4). This shows that orthogonalization is enough to guarantee that any of these shrinkage priors will always perform equally well in forecasting. However, the reader should note that this happens due to the effect of the default, noninformative priors used in this paper. For informative choices of the regularization parameters, the shrinkage penalty induced will generally be different for each of the five shrinkage priors (see the discussion in Section 3.2).

3.2. Forecasting one-year-ahead US GDP growth using the lasso

The previous subsection focused on evaluating default semi-automatic shrinkage priors using 129 variables. In practical situations, the applied macroeconomist will most probably want to focus on a few variables of interest (like inflation, an output-gap or stock prices). The previous subsection also failed to answer the question of whether other hyperparameter choices exist that could possibly make Bayesian shrinkage perform even better. Consequently, I focus here on forecasting U.S. GDP using only the simple hierarchical lasso prior, with various values being adopted for the regularization parameter λ. In particular, following De Mol et al. (2008), I use a regression model with an intercept and the 128 remaining variables as predictors (no own lags used).

The main difference from the previous subsection is that I compare four choices for λ:

- $\lambda^2 \sim \text{gamma}\,(r, \delta)$ with $r = \delta = 0.01$ (as in the benchmark case examined so far),
- $\lambda^2 \sim \text{gamma}\,(r, \delta)$ with $r = 1$ and $\delta = 0.1$,
- $\lambda^2 \sim \text{gamma}\,(r, \delta)$ with $r = 3$ and $\delta = 1$,
- λ estimated by finding the maximum marginal likelihood (MML)[9] using the Monte Carlo EM algorithm described by Park and Casella (2008).

Forecasts are generated for $h = 4$ steps ahead, and MAFE and MSFE statistics relative to the random walk

[9] This empirical Bayes procedure postulates that the hyperparameter λ used for inference should be the one that maximizes the marginal likelihood $p\,(\text{data}|\lambda)$.

Table 10
Lasso forecasts of US GDP, correlation predictors, $h = 4$.

	lasso 1	lasso 2	lasso 3	lasso 4	DFM
MAFE	0.38	0.97	0.88	0.87	0.37
MSFE	0.24	0.93	0.81	0.79	0.22
Correlation with DFM forecasts	0.90	0.27	0.49	0.48	1

Note: The lasso 1, 2, 3, 4 models are the four univariate regressions described in the text, with the estimation of λ using (a) $(r, \delta) = (0.01, 0.01)$; (b) $(r, \delta) = (1, 0.1)$; (c) $(r, \delta) = (3, 1)$; and (d) marginal maximum likelihood.

model are reported in Table 10. The benchmark prior performs the best for the US GDP, and in fact, the forecasts are highly correlated with the DFM model. Using the full sample, the estimates of the posterior median estimate of the regularization parameter λ in the four lasso models are 87.2, 33.7, 10.8 and 10.6, respectively. The prior choice $\lambda^2 \sim \text{gamma}\,(3, 1)$ gives posterior parameter estimates (and hence, forecasts) which are identical to MML estimation of λ, and this actually occurs for a wide range of choices of $r \geq 3$.

As $\lambda \rightarrow \infty$, all coefficients are penalized heavily, i.e., $\lim_{\lambda\rightarrow\infty} \boldsymbol{\beta}_{\text{lasso}} = 0$, which further implies that, in the limit, the dynamic regression model with many predictors reduces to $y_{t+h} = \boldsymbol{\alpha z}_t + \varepsilon_{t+h}$. In this case, the scale invariant prior (benchmark case) provides the largest posterior estimate of λ, which implies posterior estimates of $\boldsymbol{\beta}$ which are heavily penalized (but not exactly zero). As we allow for informative priors (cases 2 and 3), more and more variables are left unrestricted, and the results resemble the case of selection of regressors (that is, the case where some regressors are shrunk to zero and the remaining regressors are left unrestricted). For case 3, 14 coefficients are “sufficiently” different from zero, while the remaining 115 are very “low” (remember that the regressors are standardized, so it makes some sense to talk about “large” and “small” coefficients as being important or not). Nevertheless, the one-year-ahead forecasts of GDP growth are not improved when forecasting with these “14 predictors”, and hence, the benchmark case, which penalizes all predictors heavily, performs better than using an informative prior on λ. This result is robust at other forecast horizons as well (these results are not presented here). The only difference is that as the forecast horizon increases (for $h = 8$, for instance) more predictors are relevant for forecasting GDP, so that the $\lambda^2 \sim \text{gamma}\,(3, 1)$ prior leads to forecasts which are much closer to (but still dominated by) the choice $\lambda^2 \sim \text{gamma}\,(0.01, 0.01)$.

4. Concluding remarks

This paper has investigated the properties of Bayesian shrinkage using hierarchical priors. A general shrinkage representation is provided using hierarchical prior distributions, and five special cases of interest have been evaluated in forecasting using a large macroeconomic data set. A default semi-automatic approach using noninformative, near-improper priors was given special attention in this paper, but a sensitivity analysis with more informative priors for forecasting US GDP has also been carried out.

The results suggest that Bayesian shrinkage can compete favorably with factor models, although it is not straightforward to say whether one method clearly dominates the other. Both methods are efficient in reducing the dimensions of large data sets and help to achieve smaller forecast errors (especially for long-run forecasts); however, extra care has to be taken when selecting a prior for Bayesian shrinkage.

From an applied econometrician’s point of view (whether “frequentist” or “Bayesian”), the form of Bayesian shrinkage analyzed in this paper can be seen as a sensible tool which is useful for out-of-sample forecasting in the presence of many possible predictor variables (a typical every-day task for a researcher at a Federal Reserve Bank, where thousands of series are available) or when time series are short (which is part of the life of a researcher in the European Central Bank, with most Euro-Area macro series beginning in around 1995). Subsequently, this paper argues that, similarly to the very popular Bayesian model averaging and the empirical Bayes Minnesota prior for vector autoregressions mentioned in the Introduction, a “formal” (i.e., hierarchical) Bayesian treatment of the shrinkage problem should also become a standard technique for handling modern medium-to-large amounts of information.

Acknowledgments

I would like to thank a Referee, an Associate Editor, and the Editor, Rob Hyndman, for their helpful comments and suggestions, which helped to substantially improve the paper.

Appendix A. Data and transformations

All series were downloaded from St. Louis’ FRED database in December 2010, and cover the quarters Q1:1959 to Q2:2010. All series were seasonally adjusted, either taken adjusted from FRED or by applying a quarterly X11 filter based on an AR(4) model to the unadjusted series (after testing for seasonality). Some series in the database were observed only on a monthly basis, and quarterly values were computed by averaging the monthly values over the quarter (as opposed to keeping the mid-month of the quarter). All variables are transformed to be approximately stationary, and the transformation codes for each variable appear in the column ‘T’ in Table A.1.

In particular, if $w_{i,t}$ is the original untransformed series in levels, when the series is used as a predictor (RHS of Eq. (1)), the transformation codes are: (1) no transformation (levels), $x_{i,t} = w_{i,t}$; (2) first difference, $x_{i,t} = w_{i,t} -$

Table A.1
Description of series.

No	Series ID	T	F	Title
1	GDPC96	5	1	Real gross domestic product, 3 decimal
2	GDPDEF	5	1	Gross domestic product: implicit price deflator
3	PCECC96	5	1	Real personal consumption expenditures
4	PCECTPI	5	1	Personal consumption expenditures: Chain-type price index
5	GPDIC96	5	1	Real gross private domestic investment, 3 decimal
6	IMPGSC96	5	1	Real imports of goods & services, 3 decimal
7	EXPGSC96	5	1	Real exports of goods & services, 3 decimal
8	CBIC96	1	1	Real change in private inventories
9	FINSLC96	5	1	Real final sales of domestic product
10	GSAVE	5	1	Gross saving
11	GCEC96	5	1	Real government consumption expenditures & gross investment
12	SLEXPND	6	1	State & local government current expenditures
13	SLINV	6	1	State & local government gross investment
14	DPIC96	6	1	Real disposable personal income
15	PINCOME	6	1	Personal income
16	PSAVE	5	1	Personal saving
17	PRFI	6	1	Private residential fixed investment
18	PNFI	6	1	Private nonresidential fixed investment
19	PCDG	5	1	Personal consumption expenditures: Durable goods
20	PCND	5	1	Personal Consumption Expenditures: Nondurable goods
21	PCESV	5	1	Personal consumption expenditures: Services
22	GPDICTPI	6	1	Gross private domestic investment: Chain-type price index
23	WASCUR	6	1	Compensation of employees: Wages & salary accruals
24	DIVIDEND	6	1	Net corporate dividends
25	CP	6	1	Corporate profits after tax
26	CCFC	6	1	Corporate: Consumption of fixed capital
27	HOUST	4	0	Housing starts: Total: New privately owned housing units started
28	HOUST1F	4	1	Privately owned housing starts: 1-unit structures
29	HOUST5F	4	1	Privately owned housing starts: 5-unit structures or more
30	HOUSTW	4	1	Housing starts in west census region
31	HOUSTMW	4	1	Housing starts in midwest census region
32	HOUSTS	4	1	Housing starts in south census region
33	HOUSTNE	4	1	Housing starts in northeast census region
34	INDPRO	5	0	Industrial production index
35	IPCONGD	5	1	Industrial production: Consumer goods
36	IPDCONGD	5	1	Industrial production: Durable consumer goods
37	IPNCONGD	5	1	Industrial production: Nondurable consumer goods
38	IPMAT	5	1	Industrial production: Materials
39	IPDMAT	5	1	Industrial production: Durable materials
40	IPNMAT	5	1	Industrial production: Nondurable materials
41	IPBUSEQ	5	1	Industrial production: Business equipment
42	IPFINAL	5	1	Industrial production: Final products (market group)
43	UTL11	1	1	Capacity utilization: Manufacturing
44	UEMPLT5	5	1	Civilians unemployed for less than 5 weeks
45	UEMP5TO14	5	1	Civilians unemployed for 5–14 weeks
46	UEMP15T26	5	1	Civilians unemployed for 15–26 weeks
47	UEMP27OV	5	1	Civilians unemployed for 27 weeks and over
48	UNRATE	2	1	Civilian unemployment rate
49	PAYEMS	5	0	Total nonfarm payrolls: All employees
50	NDMANEMP	5	1	All employees: Nondurable goods manufacturing
51	DMANEMP	5	1	All employees: Durable goods manufacturing
52	USCONS	5	1	All employees: Construction
53	USGOOD	5	0	All employees: Goods-producing industries
54	USFIRE	5	1	All employees: Financial activities
55	USWTRADE	5	1	All employees: Wholesale trade
56	USTPU	5	1	All employees: Trade, transportation & utilities
57	USTRADE	5	1	All employees: Retail trade
58	USMINE	5	1	All employees: Natural resources & mining
59	USPBS	5	1	All employees: Professional & business services
60	USLAH	5	1	All employees: Leisure & hospitality
61	USINFO	5	1	All employees: Information services
62	USEHS	5	1	All employees: Education & health services
63	SRVPRD	5	1	All employees: Service-providing industries
64	USPRIV	5	0	All employees: Total private industries
65	USGOVT	5	1	All employees: Government
66	AHEMAN	6	1	Average hourly earnings: Manufacturing
67	AHECONS	6	1	Average hourly earnings: Construction
68	AWHMAN	5	1	Average weekly hours of production: Manufacturing
69	AWOTMAN	5	1	Average weekly hours: Overtime: manufacturing

(*continued on next page*)

Table A.1 (*continued*)

No	Series ID	T	F	Title
70	EMRATIO	5	1	Civilian employment-population ratio
71	CIVPART	5	1	Civilian participation rate
72	OPHPBS	5	1	Business sector: Output per hour of all persons
73	ULCNFB	5	1	Nonfarm business sector: Unit labor cost
74	BUSLOANS	6	1	Commercial and industrial loans at all commercial banks
75	REALLN	6	1	Real estate loans at all commercial banks
76	CONSUMER	5	1	Consumer (individual) loans at all commercial banks
77	INVEST	5	0	Total investments at all commercial banks
78	LOANS	6	0	Total loans and leases at commercial banks
79	MPRIME	2	1	Bank prime loan rate
80	GS1	2	1	1-Year treasury constant maturity rate
81	GS3	2	1	3-Year treasury constant maturity rate
82	GS5	2	1	5-Year treasury constant maturity rate
83	GS10	2	1	10-Year treasury constant maturity rate
84	FEDFUNDS	2	1	Effective federal funds rate
85	TB3MS	2	1	3-Month treasury bill: Secondary market rate
86	TB6MS	2	1	6-Month treasury bill: Secondary market rate
87	AAA	2	1	Moody's seasoned Aaa corporate bond yield
88	BAA	2	1	Moody's seasoned Baa corporate bond yield
89	M1SL	6	1	M1 money stock
90	M2SL	6	1	M2 money stock
91	CURRSL	6	1	Currency component of M1
92	DEMDEPSL	6	1	Demand deposits at commercial banks
93	SAVINGSL	6	1	Savings deposits: Total
94	TCDSL	6	0	Total checkable deposits
95	TVCKSSL	6	1	Travelers checks outstanding
96	CURRCIR	6	1	Currency in circulation
97	MZMSL	6	1	MZM money stock
98	M1V	5	1	Velocity of M1 money stock
99	M2V	5	1	Velocity of M2 money stock
100	NONREVSL	6	0	Total nonrevolving credit outstanding
101	TOTALSL	6	0	Total consumer credit outstanding
102	CPIAUCSL	6	0	Consumer price index for all urban consumers: All items
103	CPILEGSL	6	0	Consumer price index for all urban consumers: All items less energy
104	CPIULFSL	6	0	Consumer price index for all urban consumers: All items less food
105	CPIENGSL	6	1	Consumer price index for all urban consumers: Energy
106	CPIUFDSL	6	1	Consumer price index for all urban consumers: Food
107	CPIAPPSL	6	1	Consumer price index for all urban consumers: Apparel
108	CPIMEDSL	6	1	Consumer price index for all urban consumers: Medical care
109	CPITRNSL	6	1	Consumer price index for all urban consumers: Transportation
110	PPIACO	6	0	Producer price index: All commodities
111	PPIFCG	6	1	Producer price index: Finished consumer goods
112	PPIFCF	6	1	Producer price index: Finished consumer foods
113	PFCGEF	6	1	Producer price index: Finished consumer goods excluding foods
114	PPIFGS	6	1	Producer price index: Finished goods
115	PPICRM	6	1	Producer price index: Crude materials for further processing
116	PPICPE	6	1	Producer price index Finished goods: capital equipment
117	PPIITM	6	1	Producer price index: Intermediate materials: supplies & components
118	SP500	5	1	S&P 500 index
119	EXUSUK	5	1	U.S./U.K. foreign exchange rate
120	EXSZUS	5	1	Switzerland/U.S. foreign exchange rate
121	EXJPUS	5	1	Japan/U.S. foreign exchange rate
122	EXCAUS	5	1	Canada/U.S. foreign exchange rate
123	PMI	1	1	ISM manufacturing: PMI composite index
124	NAPMNOI	1	1	ISM manufacturing: New orders index
125	NAPMII	1	1	ISM manufacturing: Inventories index
126	NAPMEI	1	1	ISM manufacturing: Employment index
127	NAPMPRI	1	1	ISM manufacturing: Prices index
128	NAPMPI	1	1	ISM manufacturing: Production index
129	NAPMSDI	1	1	ISM manufacturing: Supplier deliveries index

$w_{i,t-1}$; (3) second difference, $x_{i,t} = \Delta w_{i,t} - \Delta w_{i,t-1}$; (4) logarithm, $x_{i,t} = \log w_{i,t}$; (5) first difference of logarithm, $x_{i,t} = \log w_{i,t} - \log w_{i,t-1}$; (6) second difference of logarithm, $x_{i,t} = \Delta \log w_{i,t} - \Delta \log w_{i,t-1}$.

When the series is used as the variable to be predicted (LHS of Eq. (1)), the transformation codes are: (1) no transformation (levels), $y_{i,t+h} = w_{i,t+h}$; (2) first difference, $y_{i,t+h} = w_{i,t+h} - w_{i,t}$; (3) second difference, $y_{i,t+h} = \frac{1}{h} \Delta^h w_{i,t+h} - \Delta w_{i,t}$; (4) logarithm, $y_{i,t+h} = \log w_{i,t+h}$; (5) first difference of logarithm, $y_{i,t+h} = \log w_{i,t+h} - \log w_{i,t}$; (6) second difference of logarithm, $y_{i,t+h} = \frac{1}{h} \Delta^h \log w_{i,t+h} - \Delta \log w_{i,t}$. In the transformations above, I define $\Delta w_t = w_t - w_{t-1}$ and $\Delta^h w_{t+h} = w_{t+h} - w_t$.

From the 129 series, 14 are higher level aggregates and do not add information when extracting principal components. These series are indicated with a 0 in column

Table A.2
Categories of data series based on statistical releases.

Group	Release	Number of series
1	Gross domestic product	26
2	New residential construction	7
3	G.17 Industrial production and capacity utilization	10
4	The employment situation	28
5	Productivity and costs	2
6	H.8 Assets and liabilities of commercial banks in the United States	5
7	H.15 Selected interest rates	10
8	H.6 Money stock measures	7
9	H.4.1 Factors affecting reserve balances	1
10	Money zero maturity (MZM)	1
11	Money velocity	2
12	G.19 Consumer credit	2
13	Consumer price index	8
14	Producer price index	8
15	Standard & Poor's	1
16	G.5 Foreign exchange rates	4
17	Manufacturing ISM report on business	7

'F' of the table below, and only the other 115 series are used for estimating factors.

Appendix B. Gibbs sampling algorithms using Bayesian hierarchical shrinkage priors

Note that I denote the inverse Gaussian distribution with parameters c, d as $IG\left(c, d\right)$, while the inverse gamma with parameters a, b is denoted as $\text{igamma}\left(a, b\right)$. A variable coming from the inverse gamma distribution is the reciprocal of a variable distributed as gamma, while the same is *not* true for the inverse Gaussian variate (i.e., if $z \sim IG\left(c, d\right)$, then $\left(z^{-1}\right) \sim N(c, d)$). There are many parametrizations of the gamma distribution, and the one I am using in this article is

$$\text{gamma}\left(a, b\right) \equiv f\left(w; a, b\right) = \Gamma\left(a\right) w^{a-1} b^{a} e^{-bw},$$

for a random variable w, where a, b are non-negative, real numbers, and $\Gamma\left(a\right) = \left(a-1\right)!$ is the gamma function.

In each Gibbs algorithm presented below, there is a closed form representation for the conditional of $\boldsymbol{\alpha}$ which is

$$\boldsymbol{\alpha}|\boldsymbol{\beta}, \sigma^2, \text{data} \sim N_q\left(\left(Z'Z\right)^{-1} Z'\widetilde{\boldsymbol{y}}^{\boldsymbol{\beta}}, \sigma^2\left(Z'Z\right)^{-1}\right), \qquad \text{(B.1)}$$

with $Z = \left(\boldsymbol{z}_1', \ldots, \boldsymbol{z}_T'\right)'$. That is, in the formulas of the conditional posteriors below, we need to add the sampling step in equation (B.1) above in each and every case of hierarchical prior.[10] For notational convenience, some or all of the quantities $\widetilde{\boldsymbol{y}}_{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}}$ and Ψ show up in Eq. (B.1) and in the conditional posteriors below; they are defined as $\widetilde{\boldsymbol{y}}_{\boldsymbol{\beta}} = \boldsymbol{y} - X\boldsymbol{\beta}$, $\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}} = \boldsymbol{y} - Z\boldsymbol{\alpha}$ and $\Psi = \left(\boldsymbol{y} - Z\boldsymbol{\alpha} - X\boldsymbol{\beta}\right)'\left(\boldsymbol{y} - Z\boldsymbol{\alpha} - X\boldsymbol{\beta}\right)$, respectively. Finally, in the formulas for the conditional posteriors, we have to condition on the data matrices $(\boldsymbol{y}, Z, X)$, but this is omitted for notational simplicity (to keep the formulas more compact).

B.1. Adaptive shrinkage Jeffreys prior

The priors are defined using the following hierarchy

$$\pi\left(\boldsymbol{\beta}|\tau_1^2, \ldots, \tau_p^2\right) \sim N_p\left(0, V\right)$$

$$\pi\left(\tau_j^2\right) \sim 1/\tau_j^2, \quad \text{for } j = 1, \ldots, p,$$

where $V = \text{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$. The posteriors of $\boldsymbol{\beta}$ and τ_j^2 can be obtained by sampling recursively from Eq. (B.1) and the full conditionals

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^{p} \sim N_p \times \left(\left(X'X + \sigma^2 V^{-1}\right)^{-1} X'\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}}, \sigma^2\left(X'X + \sigma^2 V^{-1}\right)^{-1}\right) \qquad \text{(B.2a)}$$

$$\frac{1}{\tau_j^2}|\boldsymbol{\alpha}, \beta_j, \sigma^2 \sim \text{gamma}\left(\frac{1}{2}, \frac{\beta_j^2}{2}\right), \quad \text{for } j = 1, \ldots, p \qquad \text{(B.2b)}$$

$$\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^{p} \sim \text{igamma}\left(\frac{T}{2}, \frac{1}{2}\Psi\right). \qquad \text{(B.2c)}$$

B.2. Adaptive shrinkage *t*-prior

The prior for this case is of the hierarchical form

$$\pi\left(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2\right) \sim N_p\left(0, \sigma^2 V\right)$$

$$\pi\left(\tau_j^2\right) \sim \text{igamma}\left(\rho, \xi\right), \quad \text{for } j = 1, \ldots, p,$$

where $V = \text{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$. Draws from the posterior can be obtained by sampling recursively from equation (B.1) and the full conditionals

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^{p} \sim N_p\left(\left(X'X + \sigma^2 V^{-1}\right)^{-1} \times X'\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}}, \sigma^2\left(X'X + \sigma^2 V^{-1}\right)^{-1}\right) \qquad \text{(B.3a)}$$

$$\frac{1}{\tau_j^2}|\boldsymbol{\alpha}, \beta_j, \sigma^2 \sim \text{gamma}\left(\rho + \frac{1}{2}, \frac{\beta_j^2}{2} + \xi\right), \quad \text{for } j = 1, \ldots, p \qquad \text{(B.3b)}$$

$$\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^{p} \sim \text{igamma}\left(\frac{T}{2}, \frac{1}{2}\Psi\right). \qquad \text{(B.3c)}$$

[10] A referee noted that it is computationally more efficient to sample $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ from a joint normal distribution, rather than sample $\boldsymbol{\alpha}|\boldsymbol{\beta}$ and then $\boldsymbol{\beta}|\boldsymbol{\alpha}$ (this is especially true when the regressors in Z are correlated to some extent with the regressors in X). The latter approach is used only for conceptual simplicity, i.e., it allows us to focus easily on expressions for $\boldsymbol{\beta}$ (which is the parameter of interest in this paper).

B.3. Hierarchical lasso

The full hierarchical representation of the lasso prior is

$$\pi\left(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2\right) \sim N_p\left(0, \sigma^2 V\right) \tag{B.4a}$$

$$\pi\left(\tau_j^2|\lambda\right) \sim \text{exponential}\left(\frac{\lambda^2}{2}\right), \quad \text{for } j = 1, \ldots, p \tag{B.4b}$$

$$\pi\left(\lambda^2\right) \sim \text{gamma}\left(r, \delta\right), \tag{B.4c}$$

where $V = \text{diag}\left\{\tau_1^2, \ldots, \tau_p^2\right\}$.

Given these priors, the posterior can be obtained by sampling recursively from Eq. (B.1) and the full conditionals

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p \sim N_p\left(\left(X'X + V^{-1}\right)^{-1} \times X'\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}}, \sigma^2\left(X'X + V^{-1}\right)^{-1}\right) \tag{B.5a}$$

$$\frac{1}{\tau_j^2}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 \sim IG\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \quad \text{for } j = 1, \ldots, p \tag{B.5b}$$

$$\lambda^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p \sim \text{gamma}\left(p + r, \frac{1}{2}\sum_{j=1}^p \tau_j^2 + \delta\right) \tag{B.5c}$$

$$\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^p \sim \text{igamma}\left(\frac{T-1}{2} + \frac{p}{2}, \frac{1}{2}\Psi + \frac{1}{2}\boldsymbol{\beta}'V^{-1}\boldsymbol{\beta}\right). \tag{B.5d}$$

B.4. Hierarchical fused lasso

The hierarchical representation of the fused lasso prior is

$$\pi\left(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2\right) \sim N_p\left(0, \sigma^2 V\right) \tag{B.6a}$$

$$\pi\left(\tau_j^2|\lambda_1\right) \sim \text{exponential}\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, \ldots, p \tag{B.6b}$$

$$\pi\left(\omega_j^2|\lambda_2\right) \sim \text{exponential}\left(\frac{\lambda_2^2}{2}\right), \quad \text{for } j = 1, \ldots, p-1 \tag{B.6c}$$

$$\pi\left(\lambda_1^2\right) \sim \text{gamma}\left(r_1, \delta_1\right) \tag{B.6d}$$

$$\pi\left(\lambda_2^2\right) \sim \text{gamma}\left(r_2, \delta_2\right), \tag{B.6e}$$

where, in this case, V is a tridiagonal matrix (see Box I).

Given these priors, the posteriors can be obtained by sampling recursively from Eq. (B.1) and the full conditionals

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\omega_j^2\right\}_{j=1}^{p-1} \sim N_p\left(\left(X'X + V_{FL}^{-1}\right)^{-1} \times X'\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}}, \sigma^2\left(X'X + V^{-1}\right)^{-1}\right) \tag{B.7a}$$

$$\frac{1}{\tau_j^2}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\omega_j^2\right\}_{j=1}^{p-1}, \sigma^2 \sim IG\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \quad \text{for } j = 1, \ldots, p \tag{B.7b}$$

$$\frac{1}{\omega_j^2}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^p, \sigma^2 \sim IG\left(\sqrt{\frac{\lambda_2^2\sigma^2}{\left(\beta_{j+1} - \beta_j\right)^2}}, \lambda^2\right), \quad \text{for } j = 1, \ldots, p-1 \tag{B.7c}$$

$$\lambda_1^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\omega_j^2\right\}_{j=1}^{p-1} \sim \text{gamma}\left(p + r_1, \frac{1}{2}\sum_{j=1}^p \tau_j^2 + \delta_1\right) \tag{B.7d}$$

$$\lambda_2^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p, \left\{\omega_j^2\right\}_{j=1}^{p-1} \sim \text{gamma}\left(p - 1 + r_2, \frac{1}{2}\sum_{j=1}^{p-1} \omega_j^2 + \delta_2\right) \tag{B.7e}$$

$$\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^p \sim \text{igamma}\left(\frac{T-1}{2} + \frac{p}{2}, \frac{1}{2}\Psi + \frac{1}{2}\boldsymbol{\beta}'V^{-1}\boldsymbol{\beta}\right). \tag{B.7f}$$

B.5. Hierarchical elastic net

For a covariance $V = \sigma^2 \times \text{diag}\left\{\left(\tau_1^{-2} + \lambda_2\right)^{-1}, \ldots, \left(\tau_p^{-2} + \lambda_2\right)^{-1}\right\}$ matrix, the hierarchical elastic net prior is

$$\pi\left(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2, \lambda_2\right) \sim N_p\left(0, V\right)$$

$$\pi\left(\tau_j^2|\lambda_1^2\right) \sim \text{exponential}\left(\frac{\lambda_1^2}{2}\right), \quad \text{for } j = 1, \ldots, p$$

$$\pi\left(\lambda_1^2\right) \sim \text{gamma}\left(r_1, \delta_1\right)$$

$$\pi\left(\lambda_2^2\right) \sim \text{gamma}\left(r_2, \delta_2\right).$$

Given these priors, the posterior can be obtained by sampling recursively from Eq. (B.1) and the full conditionals

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p \sim N_p\left(\left(X'X + V^{-1}\right)^{-1} \times X'\widetilde{\boldsymbol{y}}_{\boldsymbol{\alpha}}, \sigma^2\left(X'X + V^{-1}\right)^{-1}\right) \tag{B.8a}$$

$$\frac{1}{\tau_j^2}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 \sim IG\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda_1^2\right), \quad \text{for } j = 1, \ldots, p \tag{B.8b}$$

$$\lambda_1^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^p$$

$$V = \sigma^2 \times \begin{bmatrix} \left(\tau_1^2 + \omega_0^2 + \omega_1^2\right) & -\omega_1^2 & 0 & \cdots & 0 \\ -\omega_1^2 & \left(\tau_2^2 + \omega_1^2 + \omega_2^2\right) & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & -\omega_{p-2}^2 & 0 \\ \vdots & \ddots & -\omega_{p-2}^2 & \left(\tau_{p-1}^2 + \omega_{p-2}^2 + \omega_{p-1}^2\right) & -\omega_{p-1}^2 \\ 0 & \cdots & 0 & -\omega_{p-1}^2 & \left(\tau_p^2 + \omega_{p-1}^2 + \omega_p^2\right) \end{bmatrix}.$$

Box I.

$$\sim \text{gamma}\left(p + r_1, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta_1\right) \qquad \text{(B.8c)}$$

$$\lambda_2^2 | \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \left\{\tau_j^2\right\}_{j=1}^{p}$$

$$\sim \text{gamma}\left(\frac{p}{2} + r_2, \frac{1}{2\sigma^2}\sum_{j=1}^{p} \beta_j^2 + \delta_2\right) \qquad \text{(B.8d)}$$

$$\sigma^2 | \boldsymbol{\alpha}, \boldsymbol{\beta}, \left\{\tau_j^2\right\}_{j=1}^{p}$$

$$\sim \text{igamma}\left(\frac{T-1}{2} + \frac{p}{2}, \frac{1}{2}\Psi + \frac{1}{2}\boldsymbol{\beta}' V^{-1} \boldsymbol{\beta}\right). \qquad \text{(B.8e)}$$

References

Armagan, A., & Zaretzki, R. L. (2010). Model selection via adaptive shrinkage with *t* priors. *Computational Statistics*, *25*, 441–461.

Bai, J., & Ng, S. (2007). *Boosting diffusion indexes*. Columbia University (unpublished manuscript).

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

Fahrmeir, L., Kneib, T., & Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, *20*, 203–219.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–533.

Geweke, J. (1993). Bayesian treatment of the independent Student-*t* linear model. *Journal of Applied Econometrics*, *8*, 19–40.

Geweke, J., & Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, *26*, 216–230.

Giannone, D., Lenza, M., & Primiceri, G. E. (2012). *Prior selection for vector autoregressions*. Working Papers ECARES 2012-002, ULB - Universite Libre de Bruxelles.

Griffin, J. E., & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, *5*, 171–188.

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical mixed models. *Journal of the American Statistical Association*, *91*, 1461–1473.

Hobert, J. P., & Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, *67*, 414–430.

Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, *103*, 511–522.

Judge, G. G., & Bock, M. E. (1978). *Statistical implications of pre-test and Stein rule estimators in econometrics*. Amsterdam: North-Holland.

Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*(3), 867–886.

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

Kyung, M., Gill, J., Ghoshz, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*, 369–412.

Litterman, R. (1979). *Techniques of forecasting using vector autoregressions*. Federal Reserve Bank of Minneapolis Working Paper 115.

MacLehose, R. F., & Dunson, D. B. (2010). Bayesian semiparametric multiple shrinkage. *Biometrics*, *66*, 455–462.

Maruyama, Y., & George, E.I. (2010). *g BF: a fully Bayes factor with a generalized g-prior*. Technical Report, University of Pennsylvania. Available at http://arxiv.org/abs/0801.4410.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

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

Stock, J., & Watson, M. (2011). *Generalized shrinkage methods for forecasting using many predictors*. Unpublished manuscript. Available at http://www.princeton.edu/~mwatson/.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, *67*, 91–108.

Yi, N., & Xu, S. (2008). Bayesian lasso for quantitative trait loci mapping. *Genetics*, *179*, 1045–1055.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, *68*, 49–67.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In P. Goel, & A. Zellner (Eds.), *Bayesian inference and decision techniques*. Amsterdam: North-Holland.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.

Dimitris Korobilis is a Lecturer at the University of Glasgow. His research interests include time series analysis, Bayesian and computational econometrics, and applied macroeconomics. He has published in the *International Economic Review, Journal of Applied Econometrics*, and the *Oxford Bulletin of Economics and Statistics.*