# ECONOMIC PREDICTIONS WITH BIG DATA: THE ILLUSION OF SPARSITY

DOMENICO GIANNONE, MICHELE LENZA, AND GIORGIO E. PRIMICERI

Abstract. We compare sparse and dense representations of predictive models in macroeconomics, microeconomics and finance. To deal with a large number of possible predictors, we specify a prior that allows for both variable selection and shrinkage. The posterior distribution does not typically concentrate on a single sparse model, but on a wide set of models that often include many predictors.

## 1. Introduction

The recent availability of large datasets, combined with advances in the fields of statistics, machine learning and econometrics, have generated interest in predictive models with many possible predictors. In these cases, standard techniques such as ordinary least squares, maximum likelihood, or Bayesian inference with uninformative priors perform poorly, since the proliferation of regressors magnifies estimation uncertainty and produces inaccurate out-of-sample predictions. As a consequence, inference methods aimed at dealing with this curse of dimensionality have become increasingly popular.

As suggested by Ng (2013) and Chernozhukov et al. (2017), these methodologies can be generally divided in two broad classes. *Sparse*-modeling techniques focus on selecting a small set of explanatory variables with the highest predictive power, out of a much larger pool of regressors. For instance, best subset, forward stepwise selection, one covariate at a time multiple testing procedures, or the popular lasso belong to this class of estimators that produce sparse representations of predictive models (Beale et al., 1967, Hocking and Leslie, 1967, Draper and Smith, 1966, Chudik et al., 2018, Tibshirani, 1996, Hastie et al., 2015; see also Belloni et al., 2011a for a recent survey and examples of big data applications

*Date*: First version: March 2017. This version: July 2020.
We thank Patrick Adams, Hilde Bjørnland, Johannes Gräb, Pablo Guerron, Eric Qian and Mike West, as well as seminar and conference participants for comments and suggestions. The views expressed in this paper are those of the authors and are not necessarily reflective of views at Amazon, the European Central Bank, or the Eurosystem.

of some of these methodologies in economics). At the opposite side of the spectrum, *dense*-modeling techniques recognize that all possible explanatory variables might be important for prediction, although their individual impact might be small. This insight motivates the use of shrinkage or regularization techniques, which prevent overfitting by essentially forcing parameter estimates to be small when sample information is weak. Factor analysis or ridge regressions are standard examples of dense statistical modeling (Pearson, 1901, Spearman, 1904, Lawley and Maxwell, 1963, Tikhonov, 1963, Hoerl and Kennard, 1970, Leamer, 1973; see also Stock and Watson, 2002a,b and De Mol et al., 2008 for big data applications of these techniques in economics).

In this paper, we ask whether economic predictive problems are more likely characterized by sparsity or density. We study this question in a flexible modeling framework that encompasses both cases. In particular, we specify a so-called "spike-and-slab" prior for the coefficients of a linear predictive model, in the spirit of Mitchell and Beauchamp (1988). This prior states that regression coefficients can be non-zero with a certain probability $q$. This hyperparameter determines the degree of sparsity of the model, and we refer to it as the probability of inclusion. When a coefficient is not zero, it is modeled as a draw from a Gaussian distribution. The variance of this density is scaled by the hyperparameter $\gamma^2$, which thus controls the degree of shrinkage when a predictor is included. The higher $\gamma^2$, the higher the prior variance, the less shrinkage is performed. In sum, our model has three key ingredients. First, it allows for the possibility of sparsity by assuming that some regression coefficients may be equal to zero. Second, it shrinks the non-zero coefficients towards zero, as an alternative way to reduce estimation uncertainty and avoid overfitting for high-dimensional models. Third, it treats the degree of sparsity and shrinkage separately, as they are controlled by different hyperparameters, $q$ and $\gamma^2$. We conduct Bayesian inference on these hyperparameters, eliciting a hyperprior that is agnostic about whether to deal with the curse of dimensionality using sparsity or shrinkage.

We estimate our model on six popular datasets that have been used for predictive analyses with large information in macroeconomics, finance and microeconomics. In our macroeconomic applications, we investigate the predictability of economic activity in the US (Stock and Watson, 2002a), and the determinants of economic growth in a cross-section of countries (Barro and Lee, 1994, Belloni et al., 2011a). In finance, we study the predictability of the US equity premium (Welch and Goyal, 2008), and the factors that explain the cross-sectional

variation of US stock returns (Freyberger et al., 2017). Finally, in our microeconomic analyses, we investigate the factors behind the decline in the crime rate in a cross-section of US states (Donohue and Levitt, 2001, Belloni et al., 2014), and the determinants of rulings in the matter of government takings of private property in US judicial circuits (Chen and Yeh, 2012, Belloni et al., 2012). Notably, these six applications exhibit substantial variety, spanning time-series, cross-section and panel data, different numbers of predictors and predictors-to-observations ratios.

Our Bayesian inferential method delivers three main results. First, we characterize the marginal posterior distribution of the probability of inclusion $q$. Only in one case, the first microeconomic application, this posterior is concentrated around very low values of $q$. In all other applications, larger values of $q$ are more likely, suggesting that including more than a handful of predictors improves predictive accuracy. Second, the joint posterior distribution of $q$ and $\gamma^2$ typically exhibits a clear negative correlation: the higher the probability of including each predictor, the lower the prior variance of the non-zero coefficients. This intuitive finding highlights that larger-scale models perform well (and do not overfit) in our framework because they typically entail a higher degree of shrinkage. Third, while the appropriate degree of shrinkage and model size are quite well identified, the data are much less informative about the identity of the predictors to include or exclude from the model. Summing up, model uncertainty is pervasive, and ignoring it—as well as ignoring the evidence in favor of denser models with shrinkage—leads to an "illusion of sparsity." These findings serve as a warning against the use of sparse predictive models without critical judgement.

For an accurate interpretation of these results, it is important to appreciate the strengths and weaknesses of our approach to inference. If a prediction model with many predictors "lacks any additional structure, then there is no hope of recovering useful information about the [high-dimensional parameter] vector with limited samples" (Hastie et al., 2015, p. 290). Put differently, inference with weak assumptions is especially difficult in big data problems, and some constraints must be imposed to extract information. A widespread approach in the literature is to assume that the response variable depends on a few common factors of the predictors, which aids the recoverability of certain linear combinations of the corresponding dense parameter vector. Another popular (and polar opposite) strategy is to "bet on sparsity" (Hastie et al., 2001), by imposing that the majority of the unknown

coefficients are nearly or identically zero. When valid, these assumptions help to estimate the unknown parameters, but are not suitable to infer the degree of sparsity—the goal of this paper—since this property of the model is postulated a priori.

Our approach relaxes all sparsity and density constraints, and instead imposes some structure on the problem by making an assumption on the distribution of the non-zero coefficients. The key advantage of this strategy is that the share of non-zero coefficients is treated as unknown, and can be estimated. Another crucial benefit is that our Bayesian inferential procedure fully characterizes the uncertainty around our estimates, not only of the degree of sparsity, but also of the identity of the relevant predictors. The drawback of this approach, however, is that it might perform poorly if our parametric assumption is not a good approximation of the distribution of the non-zero coefficients. Even if we take this concern into consideration, at the very least our results show that there exist reasonable prior distributions of the non-zero regression coefficients that do not lead to sparse posteriors. More constructively, to address this robustness concern, we extend our analysis in two further directions. First, we present simulation evidence to show that our model generally recovers the true degree of sparsity. This result suggests that our "not-much-sparsity" findings do not reflect overfitting due to the inclusion of redundant predictors. Second, we demonstrate that the out-of-sample prediction performance of our model is superior to that of sparse models in our six applications.

In a related paper, Abadie and Kasy (2019) evaluate the risk of regularized estimators such as lasso, ridge and pre-testing estimators. Their analysis applies to linear regression settings like ours when the predictors are orthogonalized. In line with our approach, they emphasize the importance of selecting regularization parameters based on data-driven procedures. In addition, they conclude that pre-testing strategies perform well when the true data-generating process involves zero and non-zero regression coefficients, with the latter well separated from the former. On the contrary, ridge estimators dominate when the effects of different predictors on the dependent variable are "smoothly distributed." A strength of our model is that it encompasses these estimation strategies. They correspond to specific choices of the hyperparameters $q$ and $\gamma^2$ that can be inferred from the data, without the need to commit to dense or sparse estimators based exclusively on a-priori considerations about the data-generating process.

On a separate note, an important last point to emphasize is that the definition of sparsity is not invariant to transformations of the regressors. For example, consider a model in which only the first principal component of the explanatory variables matters for prediction. Such a model is sparse in the rotated space of the predictors corresponding to the principal components. It is instead dense in the untransformed, or "natural" space of the original regressors, since the first principal component combines all of them. This paper studies the issue of sparsity versus density in the natural space of the untransformed regressors. There are a number of reasons that motivate this focus. First, for any model, it is always possible to construct a rotated space of the predictors a posteriori, with respect to which the representation is sparse. Therefore, the question of sparsity versus density is meaningful only with respect to spaces that are chosen a priori—such as that of the original regressors or of a-priori transformations of them—and do not depend on the response variable and the design matrix. Second, our choice facilitates the comparability with the literature on lasso and variable selection, which typically assumes the existence of a sparse representation in terms of the original predictors. Third, analyzing sparsity patterns in this natural space is usually considered more interesting from an economic perspective because it may appear easier, and thus more tempting, to attach economic interpretations to models with few untransformed predictors. Before even starting to discuss whether these structural interpretations are warranted—in most cases they are not, given the predictive nature of the models—it is important to address whether the data are informative enough to clearly favor sparse models and rule out dense ones.

The rest of the paper is organized as follows. Section 2 describes the details of our prediction model. Section 3 illustrates the six economic applications. Section 4 and 5 present the main estimation results. Section 6 offers some concluding remarks.

## 2. Model

We consider the following linear model to predict a response variable $y_t$,

$$y_t = u_t'\phi + x_t'\beta + \varepsilon_t, \tag{2.1}$$

where $\varepsilon_t$ is an i.i.d. Gaussian error term with zero mean and variance equal to $\sigma^2$, and $u_t$ and $x_t$ are two vectors of regressors of dimensions $l$ and $k$ respectively, typically with $k \gg l$,

and whose variance has been normalized to one.[1] Without loss of generality, the vector $u_t$ represents the set of explanatory variables that a researcher always wants to include in the model, for instance a constant term or fixed effects in a panel regression. Therefore, the corresponding regression coefficients $\phi$ are never identically zero. Instead, the variables in $x_t$ represent possibly, but not necessarily useful predictors of $y_t$, since some elements of $\beta$ might be zero.

To capture these ideas, and address the question of whether sparse or dense representations of economic predictive models fit the data better, we specify the following prior distribution for the unknown coefficients $\left(\sigma^2, \phi, \beta\right)$,

$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$$

$$\phi \sim \text{flat}$$

$$\beta_i|\sigma^2, \gamma^2, q \sim \begin{cases} \mathcal{N}\left(0, \sigma^2\gamma^2\right) & \text{with pr. } q \\ 0 & \text{with pr. } 1-q \end{cases} \qquad i = 1, ..., k\ .$$

The priors for $\sigma^2$ and the low-dimensional parameter vector $\phi$ are rather standard, and designed to be uninformative. Instead, the elements of the vector $\beta$ are either zero, with probability $1-q$, or normally distributed with the same variance, given the standardization of the regressors. The hyperparameter $\gamma^2$ plays a crucial role since it controls the variance of this Gaussian density, and thus the degree of shrinkage when a regressor is included in the model. Without the possibility of shrinkage, the only way to improve prediction accuracy and avoid overfitting in high-dimensional models would be through variable selection. As a consequence, sparsity would emerge almost by construction.[2]

A similar way to describe our prior for $\beta$ would be to say that $\beta_i|\sigma^2, \gamma^2, q \sim \mathcal{N}\left(0, \sigma^2\gamma^2\nu_i\right)$ for $i = 1, ..., k$, with $\nu_i \sim Bernoulli\left(q\right)$. This formulation is useful because it highlights the relation between our model and some alternative specifications adopted in the literature on dimension reduction and sparse-signal detection. For example, the Bayesian ridge regression corresponds to simply setting $q = 1$. This dense model can also be interpreted as a regression

[1] The linear term in the predictors can also be interpreted as an approximation of a more general functional form. For additional flexibility, the explanatory variables can also include nonlinear transformations of the predictors, as in some of our empirical applications.

[2] As a robustness, to allow for differential degrees of shrinkage, we have also experimented with an extended version of the model in which the non-zero coefficients are drawn from a mixture of two Gaussian distributions with high and low variance, as opposed to a single one. We do not report the results based on this alternative specification, since they are similar to the baseline.

on the principal components of the $x$'s, with less shrinkage on the impact of more important principal components (Bańbura et al., 2015, Kozak et al., 2017). Therefore, this setting encompasses cases in which the bulk of the variation of $y$ and $x$ is driven by a few common factors. The Bayesian lasso, lava, horseshoe and elastic net methods can instead be obtained by replacing the Bernoulli distribution for $\nu_i$ with an exponential, a shifted exponential, a half-Cauchy, or a transformation of a truncated Gamma density, respectively (Park and Casella, 2008, Chernozhukov et al., 2017, Carvalho et al., 2010, Li and Lin, 2010). None of these alternative priors, however, admit a truly sparse representation of (2.1) with positive probability.

Our prior on $\beta$ belongs to the so-called "spike-and-slab" class, initially proposed by Mitchell and Beauchamp (1988) to perform variable selection and find sparse representations of linear regression models. Differently from them, however, the "slab" part of our prior is not a uniform density but a Gaussian, as in George and McCulloch (1993, 1997), and Ishwaran and Rao (2005). In addition, relative to most variants of the spike-and-slab prior adopted in the literature on variable selection, we treat the hyperparameters $q$ and $\gamma^2$ as unknown and evaluate their posterior distribution, along the lines of George and Foster (2000) and Liang et al. (2008). They are crucial objects of interest for our analysis of sparsity patterns.

To specify a hyperprior on $q$ and $\gamma^2$, we define the mapping $R^2\left(\gamma^2, q\right) \equiv \frac{qk\gamma^2\bar{v}_x}{qk\gamma^2\bar{v}_x+1}$, where $\bar{v}_x$ is the average sample variance of the predictors (equal to 1 in our case, given our standardization of the $x$'s). We then place the following independent priors on $q$ and $R^2$:

$$q \sim Beta\left(a, b\right)$$

$$R^2 \sim Beta\left(A, B\right).$$

The marginal prior for $q$ is a Beta distribution, with support $[0, 1]$, and shape coefficients $a$ and $b$. In our empirical applications, we will work with $a = b = 1$, which corresponds to a uniform prior. We will also experiment with prior densities that assign probability only to models with low values of $q$ and a limited number of regressors. Turning to $\gamma^2$, it is difficult to elicit a prior directly on this hyperparameter. The function $R^2\left(\gamma^2, q\right)$, instead, has the intuitive interpretation of the share of the expected sample variance of $y_t$ due to the $x_t'\beta$ term relative to the error. We model this ratio as a Beta distribution with shape coefficients $A$ and $B$, and base our inference on the uninformative case with $A = B = 1$. The appeal

of this hyperprior is that it can be used for models of possibly very different size, because it has the interpretation of a prior on the $R^2$ of the regression. Another attractive feature is that it implies a negative prior correlation between $q$ and $\gamma^2$, and is thus agnostic about whether to deal with the curse of dimensionality using sparsity or shrinkage. We will return to this point in section 4, when discussing our posterior results. They are obtained using the Markov Chain Monte Carlo algorithm for posterior evaluation detailed in appendix A.

2.1. **Simulation evidence.** Our Bayesian inferential approach relies on the exploration of the posterior distribution, which efficiently summarizes all available information about the unknown model parameters, given the observed data. Nevertheless, before moving to our empirical applications, it is interesting to conduct some Monte Carlo simulations to verify that our posterior can indeed recover the true degree of sparsity in controlled experiments.

We begin with a set of baseline simulations that obey the distributional assumptions of our model. More precisely, we generate artificial data according to (2.1), with $l = 0$, $k = 100$ and a sample size of 200. Following Belloni et al. (2011b), the predictors are drawn from a Gaussian distribution with a Toeplitz correlation matrix with $\text{corr}\left(x_{it}, x_{jt}\right) = \rho^{|i-j|}$, where we set $\rho = 0.75$. We fix $k - s$ regression coefficients to zero, and draw the remaining $s$ from a standard Normal distribution, experimenting with values of $s$ equal to 5, 10 and 100. The error term is also i.i.d. Gaussian, with variance calibrated to obtain a ratio between explained and total variance of 5, 25, and 50 percent, which is the range of degrees of predictability that characterizes the empirical applications presented in the next section. For each of these 9 designs (three values of $s$ interacted with three degrees of predictability), we simulate 100 datasets, standardize the data, and compute the posterior of the probability of inclusion.

Figure 2.1 presents a kernel approximation of the distribution of the posterior mode of $q$ across simulations. The starred dot indicates the true fraction of non-zero coefficients in each simulation design. When the degree of predictability is low, the distribution of posterior modes peaks at zero, consistent with the intuitive idea that it is difficult to detect the number of relevant regressors when their collective predictive power is very limited. If anything, a researcher would likely overstate the degree of sparsity in this case. With non-negligible predictability, however, these distributions are much more tightly concentrated around the truth, even (or, perhaps, especially) when the true data-generating process (DGP) is dense.

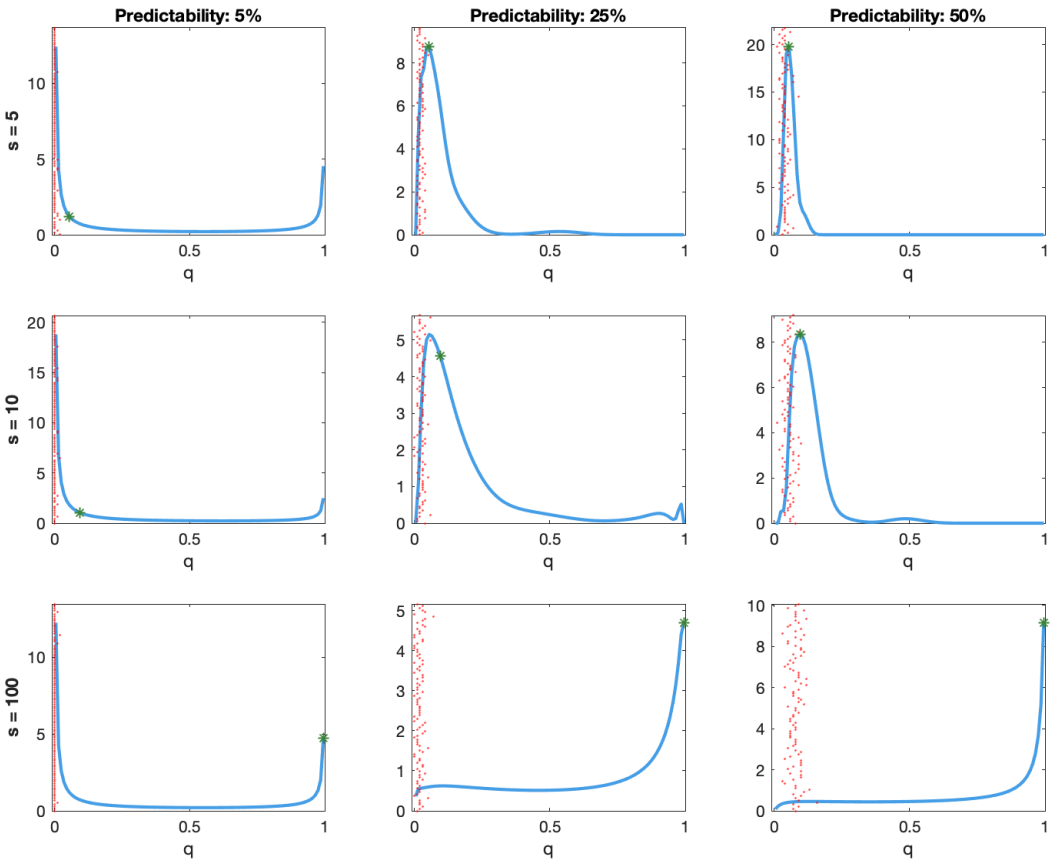


FIGURE 2.1. Baseline simulations with Gaussian and homoskedastic errors, and with non-zero coefficients drawn from a Gaussian distribution: Kernel approximation of the distribution of the posterior mode of $q$ across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by a lasso regression, with penalty parameter based on the asymptotically optimal criterion proposed by Bickel et al. (2009) and the tuning constants recommended by Belloni et al. (2011a) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

For comparison, figure 2.1 also reports the degree of sparsity estimated using a standard sparse modeling approach. More specifically, the red dots (spread along the y-axis for visibility) display the fraction of relevant predictors selected in each simulation by a lasso regression, using the asymptotically optimal value of the penalty parameter proposed by Bickel et al. (2009).[3] Notice that these lasso-based estimates of the fraction of non-zero

[3]Precisely, we report the fraction of non-zero coefficients resulting from the minimization of the penalized least square objective $\frac{1}{2T}\sum_{t=1}^{T}\left(y_t-\beta' x_t\right)^2+\lambda\sum_{i=1}^{k}|\beta_i|$. The penalization parameter is set as $\lambda=c\sqrt{\frac{\sigma^2}{T}}\,\Phi^{-1}\left(1-\frac{\zeta}{2k}\right)$, where $\Phi(\cdot)$ is the the standard normal distribution function, $1-\zeta$ is a confidence

coefficients are reasonably accurate only when the degree of predictability is high and the true $q$ is low. Instead, when the DGP has a higher number of active regressors, such as 100, lasso tends to vastly over-estimate the degree of sparsity. Intuitively, as the true DGP becomes denser, its predictive content gets spread over a higher number of predictors, and many individual coefficients become small. The lasso procedure sets these coefficients to zero, despite the fact that the collective predictive power of the associated predictors is non-negligible. On the contrary, our model can flexibly adapt to a denser DGP by increasing the degree of shrinkage. This strategy prevents overfitting by forcing parameter estimates to be small, without pushing them all the way to zero.

Our next set of simulations deviates from the homoskedasticity and Gaussianity assumptions, to provide a more challenging testing ground for our model. More specifically, we now modify the previous baseline simulations along the following three dimensions: (i) we draw the $s$ non-zero regression coefficients from a uniform distribution with mean zero and variance one, instead of a Normal; (ii) we draw the error terms from a Student-$t$ distribution with 3 degrees of freedom, instead of a Normal; (iii) we assume that these errors are heteroskedastic, instead of i.i.d., by letting the variance of $\varepsilon_t$ depend on the regressors according to $\sigma^2 \cdot \exp\left(\alpha \; x_t'\delta / \sqrt{\sum_{t=1}^{T} \left(x_t'\delta\right)^2 / T}\right)$. In this expression, $\delta$ is a $k \times 1$ parameter vector of zero (in the same positions of the zero elements of $\beta$) and non-zero elements (drawn from a standard Normal distribution). As before, $\sigma^2$ is chosen to yield the desired level of predictability. We set the parameter $\alpha$ to 4 to obtain a pronounced degree of hetroskedasticity, with the variance of $\log\left(\text{var}\left(\varepsilon\right)\right)$ equal to 4.

Figure 2.2 presents the outcome of this second simulation exercise, showing that the model is able to detect the true level of sparsity as well as in the baseline case, even in this considerably more challenging environment. In appendix B we show similarly successful recovery patterns for some additional simulation designs in which the errors are heteroskedastic and non-Gaussian, and the regression coefficients are drawn from a Laplace or mixtures of Gaussian distributions, instead of a uniform.

These results are comforting, as they show that our model is able to detect the true level of sparsity even if its parametric assumptions are substantially different from those of the DGP. What type of extreme assumptions about the DGP could then undermine the

level and $c$ is an arbitrary constant. Following Belloni et al. (2011a), we set $\zeta = 0.05$ and $c = 1.1$. The variance of the residuals $\sigma^2$ is set at the true value, which is known in our controlled experiment.

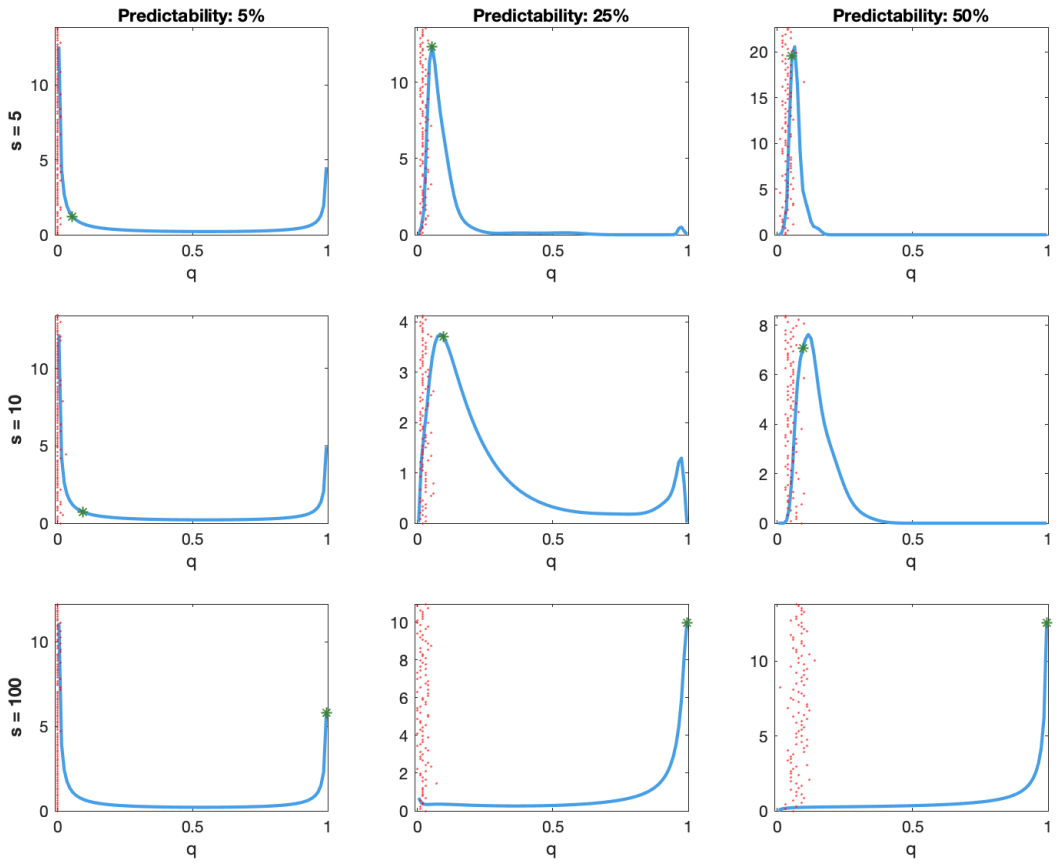


FIGURE 2.2. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a uniform distribution: Kernel approximation of the distribution of the posterior mode of $q$ across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by a lasso regression, with penalty parameter based on the asymptotically optimal criterion proposed by Bickel et al. (2009) and the tuning constants recommended by Belloni et al. (2011a) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

performance of the method? We explore this question in our last two sets of simulations, which are designed to "test the boundaries" of our model. Given the focus of this paper, we are particularly interested in uncovering situations in which the true DGP is sparse, but the posterior mode of $q$ is likely to erroneously point towards density.

For this reason, our next experiment analyzes the case in which the true DGP is not exactly sparse, but only approximately so, in the sense of Belloni et al. (2011a). Specifically, we repeat the baseline simulations of figure 2.1 with $s = 5$. However, instead of setting

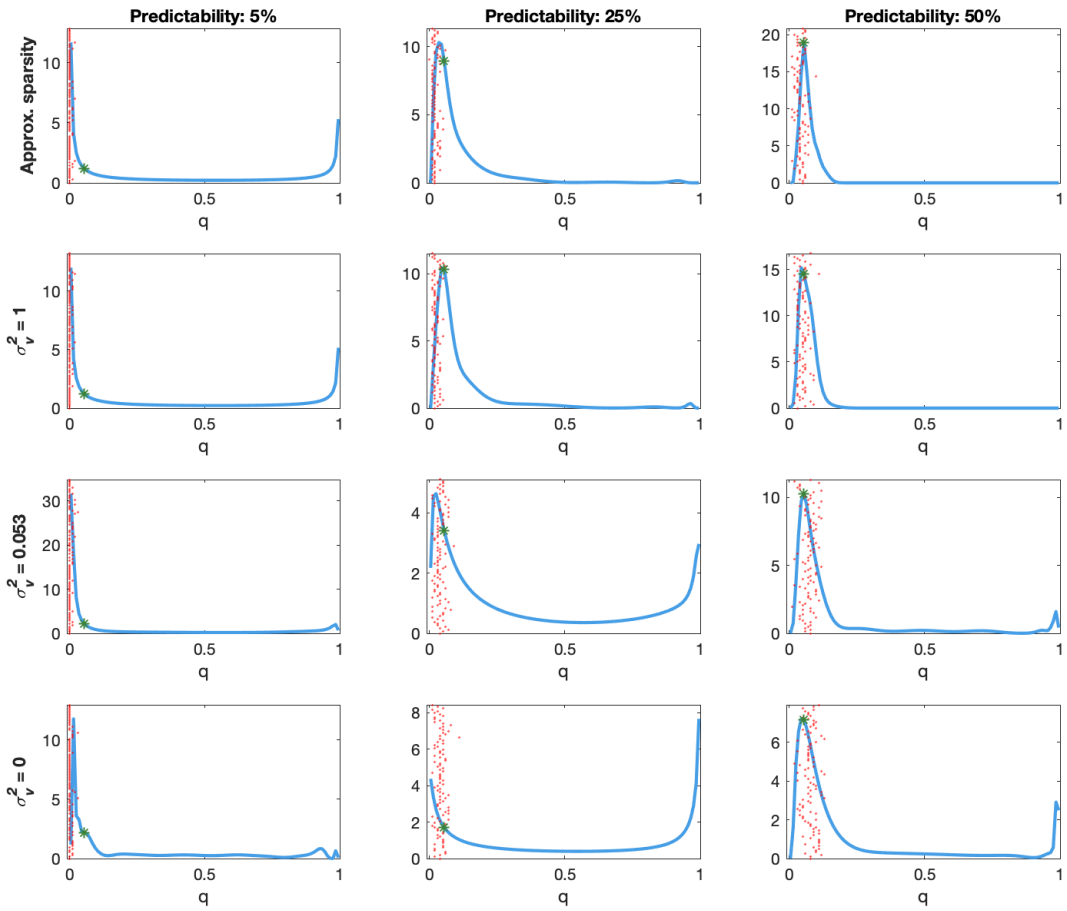


FIGURE 2.3. Simulations with approximate sparsity (first row) and a factor structure for the predictors (second, third and fourth rows): Kernel approximation of the distribution of the posterior mode of $q$ across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by a lasso regression, with penalty parameter based on the asymptotically optimal criterion proposed by Bickel et al. (2009) and the tuning constants recommended by Belloni et al. (2011a) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

the remaining $k - s$ regression coefficients to zero, we draw them from a standard Normal distribution, and then re-scale them so that the combined effect of the corresponding $k - s$ regressors on the response variable has a variance equal to $\sigma^2 \frac{s}{T}$. As evident from the first row of figure 2.3, the model is still able to detect the true level of sparsity quite well, even though sparsity is now contaminated by noise.

Our final experiment captures situations in which a few active predictors are highly correlated with many inactive ones. In these cases, a linear combination of the latter might be able to proxy for the former, and our posterior analysis might then point towards density, even if the DGP is sparse. Unfortunately, implementing this intuition in a simulation is not as simple as assuming a very high correlation among all the predictors, for example $\text{corr}\,(x_{it}, x_{jt}) = 0.9$ for $i \neq j$, instead of the Toeplitz correlation matrix of our baseline simulations. The reason is that a widespread high correlation among all regressors also increases the extent to which the inactive regressors are collinear with each other, making it harder for them to span the space of the true regressors.[4] Hence, we need a more flexible framework that allows us to boost the correlation between the active and the inactive predictors, while keeping the correlation among the inactive predictors at low values. This is accomplished by generating the regressors using the factor structure

$$x_t^{ac} = f_t + v_t$$

$$x_t^{in} = \Lambda f_t + w_t,$$

where $x_t^{ac}$ are the $s$ active predictors, $x_t^{in}$ are the $k-s$ inactive ones, and $x_t = \left[x_t^{ac\prime}\ x_t^{in\prime}\right]'$. In these expressions, $f_t$ is an $s$-dimensional vector of common factors and $\Lambda$ is a $(k-s)\times s$ matrix of loadings, all drawn from standard Gaussian distributions. The errors $v_t$ and $w_t$ are also Normal. We calibrate the variance of $w_t$ so that the common factors explain 50 percent of the variance of $x_t^{in}$, on average. As for the variance of $v_t$, we experiment with $\sigma_v^2$ equal to 1, 0.053, and 0, which imply that the ratio between the variance of the common factors and that of $x_t^{ac}$ is 50, 95 or 100 percent. After generating the $x$'s as just described, the rest of the simulation is identical to the baseline.

The second, third and fourth row of figure 2.3 present the outcome of this last set of simulations. When $\sigma_v^2$ is high and $x_t^{ac}$ are imperfect proxies of the common factors (as imperfect as $x_t^{in}$), the model is still able to recover the true degree of sparsity. The performance of the model starts to deteriorate only when $\sigma_v^2$ is very low or zero, corresponding to the admittedly extreme circumstance in which the variables $x_t^{ac}$ are almost or exactly equal to the common factors. In this case, everything continues to work well if the degree of predictability is 50 percent. If it is equal to 25 percent, however, the posterior distribution

[4] In fact, when we simulate artificial data with $\text{corr}\,(x_{it}, x_{jt}) = 0.9$ for $i \neq j$, instead of the Toeplitz correlation matrix, the model continues to perform very well.

often peaks around high values of $q$, suggesting that in many simulations a linear combination of all inactive predictors can span the space of the true ones. This said, we argue that it is unclear whether this result signals a failure of the model to detect the true level of sparsity, given that these last simulations are exactly designed so that a dense model is a good approximation of the sparse one.

## 3. Economic Applications

We estimate the previous model on six popular "big datasets" that have been used for predictive analyses in macroeconomics, finance and microeconomics. In our macroeconomic applications, we investigate the predictability of economic activity in the US (macro 1) and the determinants of economic growth in a cross-section of countries (macro 2). In finance, we study the predictability of the US equity premium (finance 1) and the factors that explain the cross-sectional variation in expected US stock returns (finance 2). Finally, in our microeconomic applications, we investigate the effects of legalized abortion on crime in a cross-section of US states (micro 1) and the determinants of rulings in the matter of government takings of private property in US judicial circuits (micro 2). Many of these applications are true "classics" in their respective literatures. Moreover, they collectively exhibit substantial variety, spanning time-series, cross-section and panel data, different numbers of predictors and predictors-to-observations ratios. Table 1 summarizes the data used in the analysis. A more detailed description is provided in the next subsections.

3.1. **Macro 1: Macroeconomic forecasting using many predictors.** In this application, we study the importance of large information to forecast US economic activity, an issue investigated by a large body of time-series research in the last decade. We use a popular large dataset originally developed for macroeconomic predictions with principal components by Stock and Watson (2002a,b), and extensively used to assess the forecasting performance of alternative big-data methodologies. The variable to predict is the monthly growth rate of US industrial production, and the dataset consists of 130 possible predictors, including various monthly macroeconomic indicators, such as measures of output, income, consumption, orders, surveys, labor market variables, house prices, consumer and producer prices, money, credit and asset prices. The constant term is always included as a regressor. The sample ranges from February 1960 to December 2014, and all the data have been transformed to obtain stationarity, as in the work of Stock and Watson. The version of

| | Dependent variable | Possible predictors | Sample |
|---|---|---|---|
| **Macro 1** | Monthly growth rate of US industrial production | 130 lagged macroeconomic indicators | 659 monthly time-series observations, from February 1960 to December 2014 |
| **Macro 2** | Average growth rate of GDP over the sample 1960-1985 | 60 socio-economic, institutional and geographical characteristics, measured at pre-1960s value | 90 cross-sectional country observations |
| **Finance 1** | US equity premium (S&P 500) | 16 lagged financial and macroeconomic indicators | 68 annual time-series observations, from 1948 to 2015 |
| **Finance 2** | Stock returns of US firms | 144 dummies classifying stock as very low, low, high or very high in terms of 36 lagged characteristics | 1400k panel observations for an average of 2250 stocks over a span of 624 months, from January 1963 to May 2014 |
| **Micro 1** | Per-capita crime (murder) rates | Effective abortion rate and 284 controls including possible covariate of crime and their transformations | 576 panel observations for 48 US states over a span of 144 months, from 1986 to 1997 |
| **Micro 2** | Number of pro-plaintiff eminent domain decisions in a specific circuit and in a specific year | Characteristics of judicial panels capturing aspects related to gender, race, religion, political affiliation, education and professional history of the judges, together with some interactions among the latter, for a total of 138 regressors | 312 panel observations for 12 circuits over a span of 26 years, from 1979 to 2004 |

TABLE 1. Description of the datasets.

the dataset that we use is available at FRED-MD, and is regularly updated through the Federal Reserve Economic Data (FRED), a database maintained by the Research division of the Federal Reserve Bank of St. Louis (McCracken and Ng, 2016).

3.2. **Macro 2: The determinants of economic growth in a cross-section of countries.** The seminal paper by Barro (1991) initiated a debate on the cross-country determinants of long-term economic growth. Since then, the literature has proposed a wide range of possible predictors of long-term growth, most of which have been collected in the dataset constructed by Barro and Lee (1994). As in Belloni et al. (2011a), we use this dataset

to explain the average growth rate of GDP between 1960 and 1985 across countries. The database includes data for 90 countries and 60 potential predictors, corresponding to the pre-1960 value of several measures of socio-economic, institutional and geographical characteristics. The constant term and the logarithm of a country's GDP in 1960 are always included as regressors in the model.[5]

3.3. **Finance 1: Equity premium prediction.** Following a large body of empirical work, in our first finance application we study the predictability of US aggregate stock returns, using the database described in Welch and Goyal (2008). More specifically, the dependent variable is the US equity premium, defined as the difference between the return on the S&P 500 index and the 1-month Treasury bill rate. As possible predictors, the dataset includes sixteen lagged variables deemed as relevant in previous studies, such as stock characteristics (the dividend-price ratio, the dividend yield, the earning-price ratio, the dividend-payout ratio, the stock variance, the book-to-market ratio for the Dow Jones Industrial Average, the net equity expansion and the percent equity issuing), interest rate related measures (the Treasury bill, the long-term yield, the long-term return, the term spread, the default-yield spread and the defaults-return spread) and some macroeconomic indicators (inflation and the investment-to-capital ratio). The constant term is always included as a regressor. The data are annual, and the sample ranges from 1948 to 2015.[6]

3.4. **Finance 2: Explaining the cross section of expected returns.** Despite the simple characterization of equity returns provided by the workhorse CAPM model, the empirical finance literature has discovered many factors that can explain the cross-section of expected asset returns. The recent survey of Harvey et al. (2016) identifies about 300 of these factors. Following this tradition, in this application we study the predictability of the cross-section of US stock returns, based on the dataset of Freyberger et al. (2017).[7] Our dependent variable is the monthly stock return of firms incorporated in the US and trading on NYSE, Amex and Nasdaq, from January 1963 to May 2014, which results in about 1,400k observations. The set of potential regressors are constructed using (the lagged value of) 36 firm and stock characteristics, such as market capitalization, the return on assets

[5]We have downloaded the dataset from the replication material of Belloni et al. (2011a), who consider exactly the same application.

[6]We use an updated version of the database downloaded from the webpage of Amit Goyal.

[7]We thank Joachim Freyberger, Andreas Neuhierl and Michael Weber for sharing the database used in their paper.

and equity, the book-to-market ratio, the price-dividend ratio, etc. Inspired by the flexible nonparametric approach of Freyberger et al. (2017), for each of these characteristics we create four dummy variables that take the value of one if the firm belongs to the first, second, fourth or fifth quintile of the distribution within each month, respectively.[8] This results in 144 possible regressors, plus a constant term (always included as a regressor).

3.5. **Micro 1: Understanding the decline in crime rates in US states in the 1990s.** Using US state-level data, Donohue and Levitt (2001) find a strong relationship between the legalization of abortion following the Roe vs Wade trial in 1973, and the subsequent decrease in crime rates. Their dependent variable is the change in log per-capita murder rates between 1986 and 1997 across US states. This variable is regressed on a measure of the effective abortion rate (which is always included as a predictor, along with 12 month dummy variables) and a set of controls. The latter capture other possible factors contributing to the behavior of crime rates, such as the number of police officers per 1000 residents, the number of prisoners per 1000 residents, personal income per capita, the unemployment rate, the level of public assistance payments to families with dependent children, beer consumption per capita, and a variable capturing the shall-issue concealed carry laws. In addition, as in Belloni et al. (2014), we expand the set of original controls of Donohue and Levitt (2001), by including these variables in levels, in differences, in squared-differences, their cross-products, their initial conditions and their interaction with linear and squared time trends. This extended database includes 284 variables, each with 576 observations relating to 48 states for 12 years.[9]

3.6. **Micro 2: The determinants of government takings of private property in US judicial circuits.** Chen and Yeh (2012) investigate the economic impact of eminent domain, i.e. the right of the US government to expropriate private property for public use. To address the possible endogeneity problem, they propose to instrument judicial decisions on eminent domain using the characteristics of randomly assigned appellate courts judges. We follow Belloni et al. (2012) and estimate the first stage of this instrumental-variable model, by regressing the number of pro-plaintiff appellate decisions in takings law rulings

[8]For collinearity reasons, we exclude the dummy variable that is equal to one if the firm belongs to the third quintile.

[9]We downloaded the data from the replication material of Belloni et al. (2014), who consider exactly the same application.

from 1979 to 2004 across circuits on a set of characteristics of the judicial panels such as gender, race, religion, political affiliation, education and professional history of the judges. As in Belloni et al. (2012), we augment the original set of instruments with many interaction variables, resulting into 138 possible regressors. As in their work, the regression always includes the constant term, a set of year and circuit dummy variables, and circuit-specific time trends. The sample size (circuit/year units) consists of 312 observations.[10]

## 4. Exploring the Posterior

In this section, we discuss some properties of the posterior distribution of our model, estimated using the six datasets illustrated in the previous section. The results we report are based on a uniform prior on $q$ and $R^2$, i.e. the probability of inclusion and the share of the expected sample variance of $y_t$ explained by the predictors. We will also explore the implications of priors concentrated on low values of $q$.

### 4.1. Positive correlation between probability of inclusion and degree of shrinkage.

Our inference method allows us to characterize the joint distribution of the hyperparameters $q$ and $\gamma^2$, i.e. the probability of inclusion and the prior variance of the coefficients of the included predictors. The left panels of figures 4.1 and 4.2 summarize the shape of the prior of these two hyperparameters in our six empirical applications, with lighter areas corresponding to higher density regions.[11] We present the joint density of $q$ and $log(\gamma)$, instead of $q$ and $\gamma^2$, to interpret the horizontal axis more easily in terms of percent deviations. As we noted in section 2, our flat prior on $q$ and $R^2$ implies a negative correlation between $q$ and $log(\gamma)$, reflecting the sensible prior belief that sparsity and shrinkage are substitutes when it comes to deal with the curse of dimensionality.

The right panels of figures 4.1 and 4.2 show the posteriors of $q$ and $log(\gamma)$. These densities are typically much more concentrated than the corresponding prior, exhibiting an even sharper negative correlation: the lower (higher) the probability of including a predictor and the overall model size, the higher (lower) the prior variance of the coefficients of the predictors included in the model. In other words, larger-scale models need more shrinkage

[10] We have downloaded the dataset from the replication material of Belloni et al. (2012), who consider exactly the same application.

[11] The darkness of these plots is also adjusted to correctly capture the scale of the prior relative to the corresponding posterior.

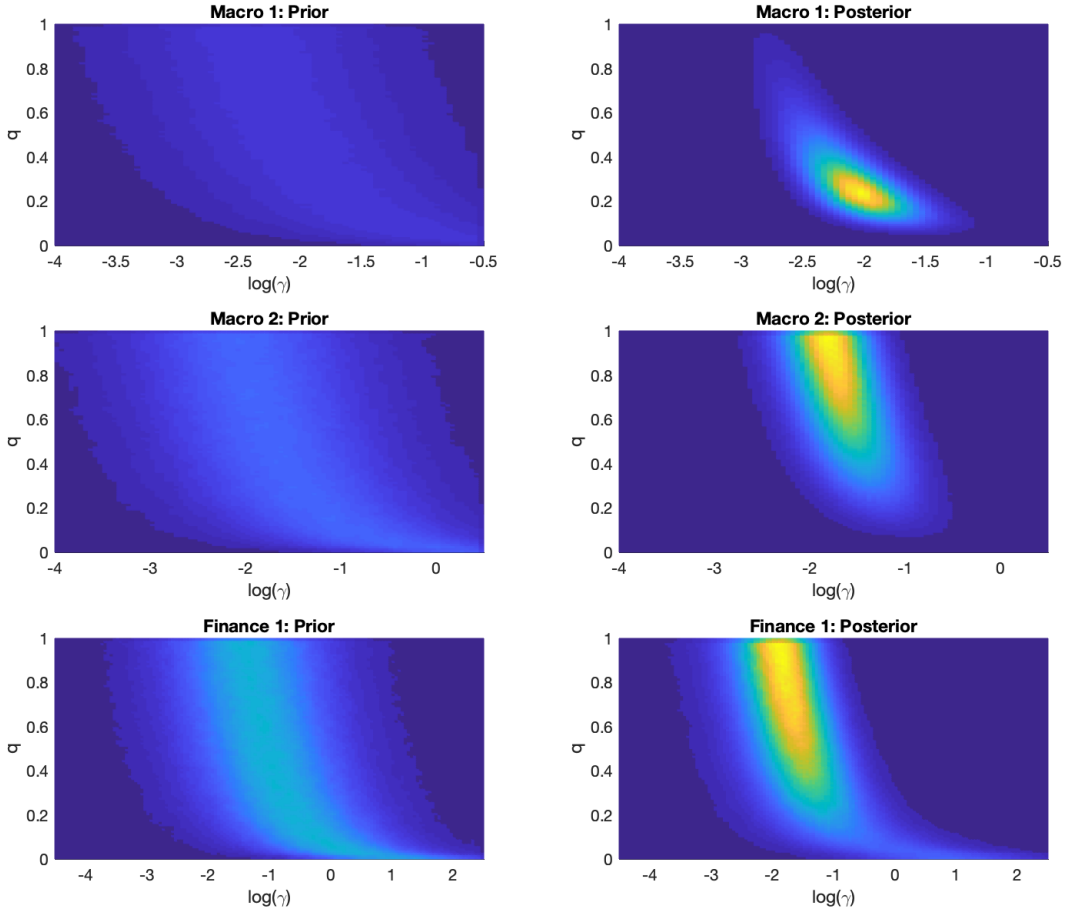


FIGURE 4.1. Joint prior and posterior densities of $q$ and $log\,(\gamma)$ in the macro-1, macro-2 and finance-1 applications (best viewed in color).

to fit the data well, while models with a low degree of shrinkage require the selection of fewer explanatory variables.

While this result should not be particularly surprising, its important implication is that variable selection techniques that do not explicitly allow for shrinkage might artificially recover sparse model representations simply as a device to reduce estimation error. Our findings indicate that these extreme strategies might perhaps be appropriate only for our micro-1 application, given that its posterior in figure 4.2 is tightly concentrated around extremely low values of $q$. More generally, however, our results suggest that the best predictive models are those that optimally combine probability of inclusion and shrinkage.

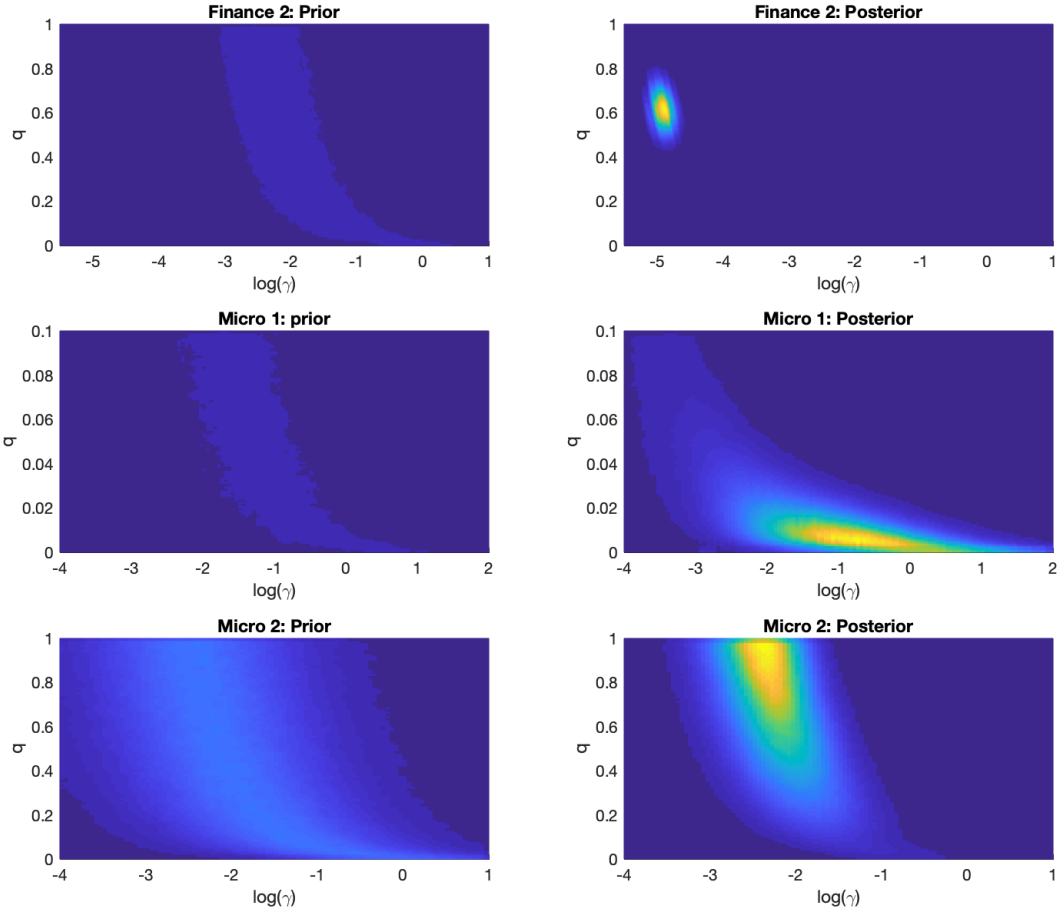


FIGURE 4.2. Joint prior and posterior densities of $q$ and $log\,(\gamma)$ in the finance-2, micro-1 and micro-2 applications (best viewed in color).

4.2. **Probability of inclusion and degree of sparsity.** What is then the appropriate probability of inclusion, considering that models with different sizes require different shrinkage? To answer this question, figure 4.3 plots the marginal posterior of $q$, obtained by integrating out $\gamma^2$ from the joint posterior distribution of figures 4.1 and 4.2. Notice that the densities in figure 4.3 behave quite differently across applications. For example, the finance-1 data seem to contain little information about model size, since the posterior of $q$ peaks at 1, but deviates little from its uniform prior. The macro-2 and micro-2 applications more strongly favor dense models with the full set of predictors. At the opposite extreme, micro 1 is the only application in which the posterior density is concentrated on very low values of $q$, suggesting that the model is likely sparse. Macro 1 and finance 2, instead,

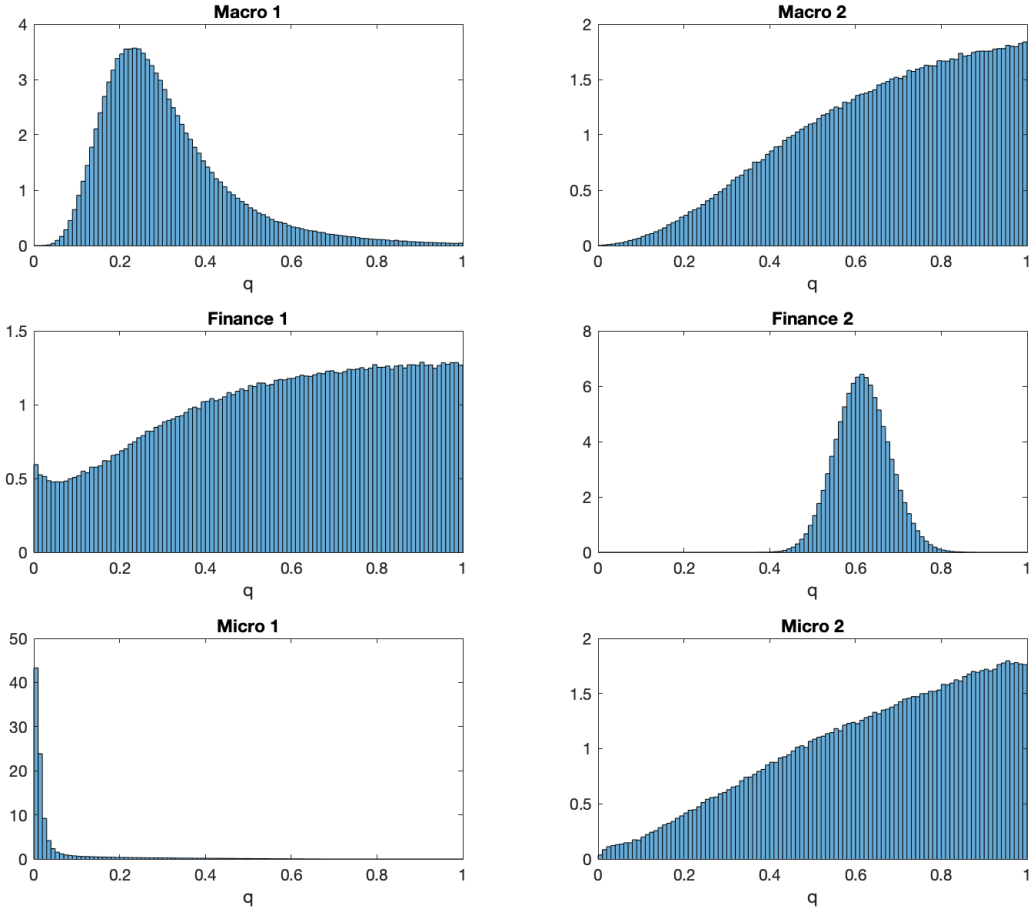


FIGURE 4.3. Posterior density of $q$.

represent intermediate cases, in which the posterior of $q$ is nicely shaped and peaks at an interior point of the $[0, 1]$ interval.

4.3. **Model uncertainty and patterns of sparsity.** The previous subsection has presented evidence about the *share* of relevant predictors, i.e. the degree of sparsity in our six applications. Given these results, we now ask whether the *identity* of these relevant predictors—i.e. the pattern of sparsity—is well identified, especially in macro 1, finance 2 and micro 1, for which the posterior of $q$ does not peak at 1. We will see that this is the case only in our micro-1 application. For macro 1 and finance 2, instead, there is a lot of uncertainty about the identity of the relevant predictors.

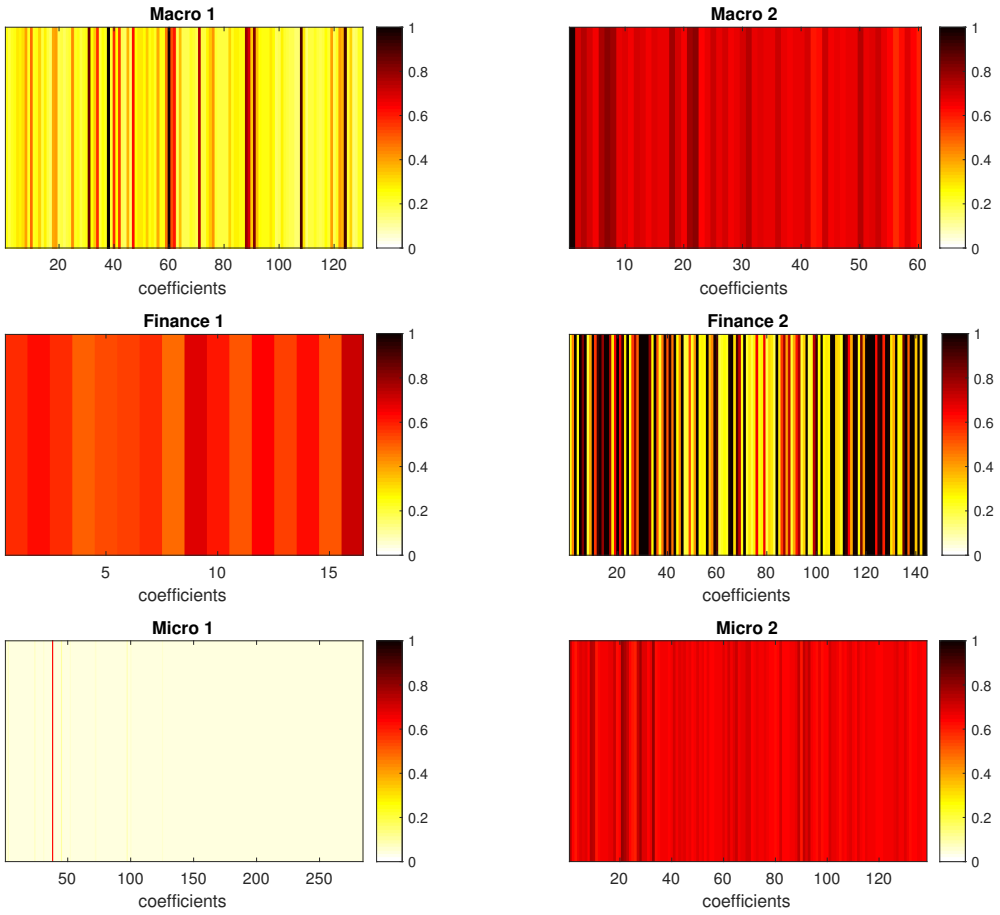


FIGURE 4.4. Heat map of the probabilities of inclusion of each predictor (best viewed in color).

To illustrate this point, figure 4.4 plots the posterior probabilities of inclusion of each predictor. In the "heat maps" of this figure, each vertical stripe corresponds to a possible predictor, and darker shades denote higher probabilities of inclusion. Notice that the probability of inclusion of a single regressor might deviate considerably from $q$, although the average probability of inclusion across regressors should be consistent with the posterior of $q$.

The most straightforward subplot to interpret is the one corresponding to the micro-1 application. This is a model with a clear pattern of sparsity, in which the 39th regressor (income squared) is selected 65 percent of the times. The 46th regressor is also sometimes

selected, about 10 percent of the times, although this is more difficult to see from the plot. All other predictors are included in the model much more rarely.

The important message of figure 4.4, however, is that the remaining five applications do not exhibit a distinct pattern of sparsity, in the sense that none of the predictors appear to be systematically excluded. This finding was probably expected for macro 2, finance 1 and micro 2, since the posterior of $q$ peaks around very high values in these three applications. The absence of clear sparsity patterns, however, should be more surprising when the posterior of $q$ has most of its mass on lower values. For example, let us consider the case of macro 1, in which the best fitting models are those with $q$ around 0.23, according to figure 4.3. This value of $q$, however, does not necessarily imply that the most accurate model includes 30 specific predictors (23 percent of the 130 possible regressors) and excludes all the others. If this were the case, the first panel of figure 4.4 would show many near-white stripes corresponding to the predictors that are systematically excluded. Instead, there seems to be a lot of uncertainty about whether certain predictors should be included in the model, which results into their selection only in a subset of the posterior draws. These findings may also reflect a substantial degree of collinearity among many predictors that carry similar information, hence complicating the task of structure discovery.

In sum, according to our results, model uncertainty is pervasive and the best prediction is obtained as a weighted average of several models with different sets of regressors. These findings, in turn, rationalize the empirical success of model averaging techniques and, more generally, ensemble machine learning methods such as boosting, bagging and random forests. Examples of applications of these tools to various fields of economics include Wright (2009), Faust et al. (2013), Fernandez et al. (2001), Sala-I-Martin et al. (2004), Cremers (2002), Avramov (2002), Inoue and Kilian (2008), Bai and Ng (2009), Rapach and Strauss (2010), Ng (2014), Jin et al. (2014), Varian (2014), Wager and Athey (2018) and Athey et al. (2019), among others.

How can we reconcile these results with the large literature on sparse signal detection, which is also gaining popularity in economics? The short answer is that, to guarantee the recoverability of the "true" model, this literature typically needs to rule out a priori the possibility that such true model is high-dimensional, even when the number of possible predictors is large. This is achieved by either adopting explicit priors favoring low-dimensional models (e.g. Castillo et al., 2015), or formal assumptions that the size of the true model is

small relative to $k$ (e.g. Bickel et al., 2009 and the subsequent related literature on penalized regressions). Intuitively, detecting the relevant predictors might become easier when only considering models with a few of them, for a given sample size.

To verify this intuition, figure 4.5 visualizes the probability of inclusion of each predictor, conditional on all possible values of $q$. The horizontal line denotes the posterior mode of $q$ in each application.[12] Observe that high-density values of $q$—values close to the mode—are mostly associated with non-white regions in the heat map (except for micro 1), confirming the main takeaway of figure 4.4 about the absence of clear sparsity patterns. When $q$ is very low, instead, some regressors are excluded more systematically and others emerge as relatively more important. Figure 4.5 makes clear that a researcher with a dogmatic prior that $q$ must be low would artificially detect somewhat clearer patterns of sparsity. It is important to remark, however, that even in these circumstances model uncertainty does not completely vanish. Moreover, even a relatively small degree of model uncertainty can have sizable consequences when models are low-dimensional, since they do not overlap as much as large models and thus may contain useful idiosyncratic information. As we will see in the next section, ignoring this uncertainty can be costly in terms of predictive accuracy.

## 5. Implications for Out-of-Sample Predictive Accuracy

The results of the previous section show that (i) the posterior of $q$ does not generally concentrate on low values, and (ii) model uncertainty is substantial. One implication of these findings is that sparsity-based methods—yielding low-dimensional models without uncertainty on the identity of the predictors—are likely to lead to predictive losses. In this section, we document that this is indeed the case. We begin by showing that constraining the model space to only include small models is generally costly in terms of predictive ability. We then demonstrate that going one step further and also ignoring model uncertainty can be even more detrimental.

5.1. **Predictive accuracy and model size.** For a first pass at understanding whether low-dimensional models entail predictive losses, notice that the posterior of model size can be directly interpreted as a measure of out-of-sample predictive accuracy. To see this, let

[12]These heat maps are truncated in finance 2 and micro 1 because the posterior of $q$ is very concentrated in these applications, and there are essentially no posterior draws of $q$ below or above those thresholds.

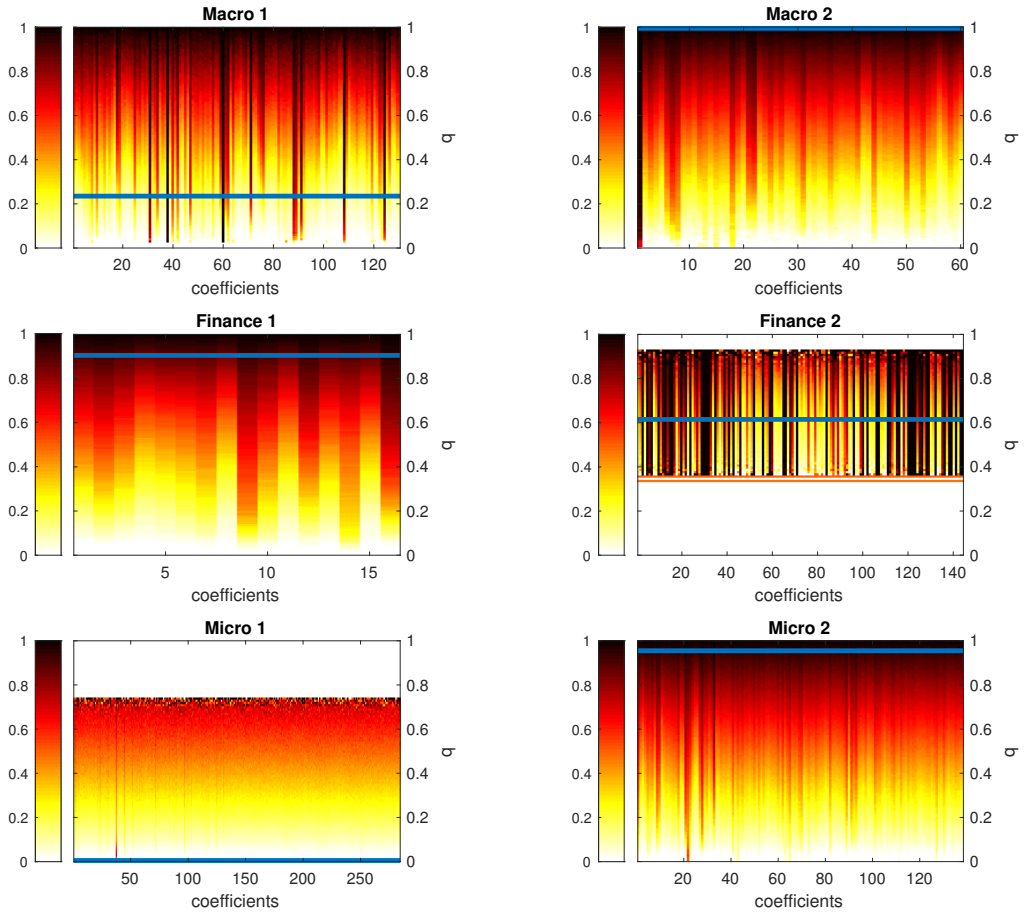


FIGURE 4.5. Heat map of the probabilities of inclusion of each predictor, conditional on $q$. The horizontal line denotes the posterior mode os $q$ (best viewed in color).

$s$ denote the size of the model, i.e. the number of active predictors. Our flat prior on $q$ implies a flat prior on $s$, and its posterior is thus proportional to the likelihood,

$$p(s|y) \propto p(y|s) = \prod_{t=1}^{T} p\left(y_t|y^{t-1}, s\right), \tag{5.1}$$

where the equality follows from the usual decomposition of a joint density into the product of marginal and conditional densities, and we are omitting the regressors $u$ and $x$ from the conditioning sets to streamline the notation. Expression (5.1) makes clear that the posterior

of $s$ is proportional to a product of predictive scores. As a consequence, the values of $s$ with higher posterior density are also those associated with better out-of-sample predictions.

Figure 5.1 quantifies the variation in predictive accuracy across models with different $s$, by plotting the function $\log p(y|s) - \log p(y|s^*)$ in our six economic applications (solid line). This expression corresponds to the log-predictive score of a model with $s$ predictors, relative to the model with $s^*$ predictors, where $s^*$ denotes the posterior mode of $s$. For instance, in our macro-1 application, values close to the actual realizations of $y$ are more likely according to a model with $s \approx 30$ relative to those with very low or high $s$. As for macro 2, the best predictive model is the dense one with $s = k$, and the top-right panel of figure 5.1 summarizes the deterioration in the log-predictive score when $s$ declines. The dense model is also marginally preferred for finance 1, but figure 5.1 makes clear that predictive accuracy varies little across model sizes for this application, due to the very low degree of predictability and the limited number of observations. The remaining panels in figure 5.1 can be interpreted in a similar way. The overall conclusion is that, with the exception of micro 1, small models with a handful of predictors are associated with lower predictive accuracy.

Although theoretically elegant, measuring relative predictive accuracy using the posterior of $s$ is not fully satisfactory, because this measure can only be used to rank models with different fixed values of $s$. For example, it does not allow the evaluation of the relative predictive accuracy of a model where the researcher conducts inference on $s$ in real time, or of models with specific sets of regressors. To broaden the scope of the analysis, we also conduct a fully-fledged out-of-sample forecasting exercise, whose implementation details are described in appendix C. In this exercise, we re-estimate the model on many training samples, obtained as subsets of the full sample. We then evaluate the predictive performance of our model on many corresponding test samples, comparing it to that obtained by restricting the model space in a variety of informative ways.

The horizontal lines in figure 5.1 represent the resulting log-predictive scores of four versions of the model: SS-bma, which is our full model that combines all the possible individual models, weighted by their posterior probability (dashed line);[13] SS-bma-5 and SS-bma-10, which restrict the model space to the combinations of individual models with

[13]SS-bma is an abbreviation of a Bayesian model averaging (bma) strategy, with posterior weights obtained using our spike-and-slab (SS) prior.

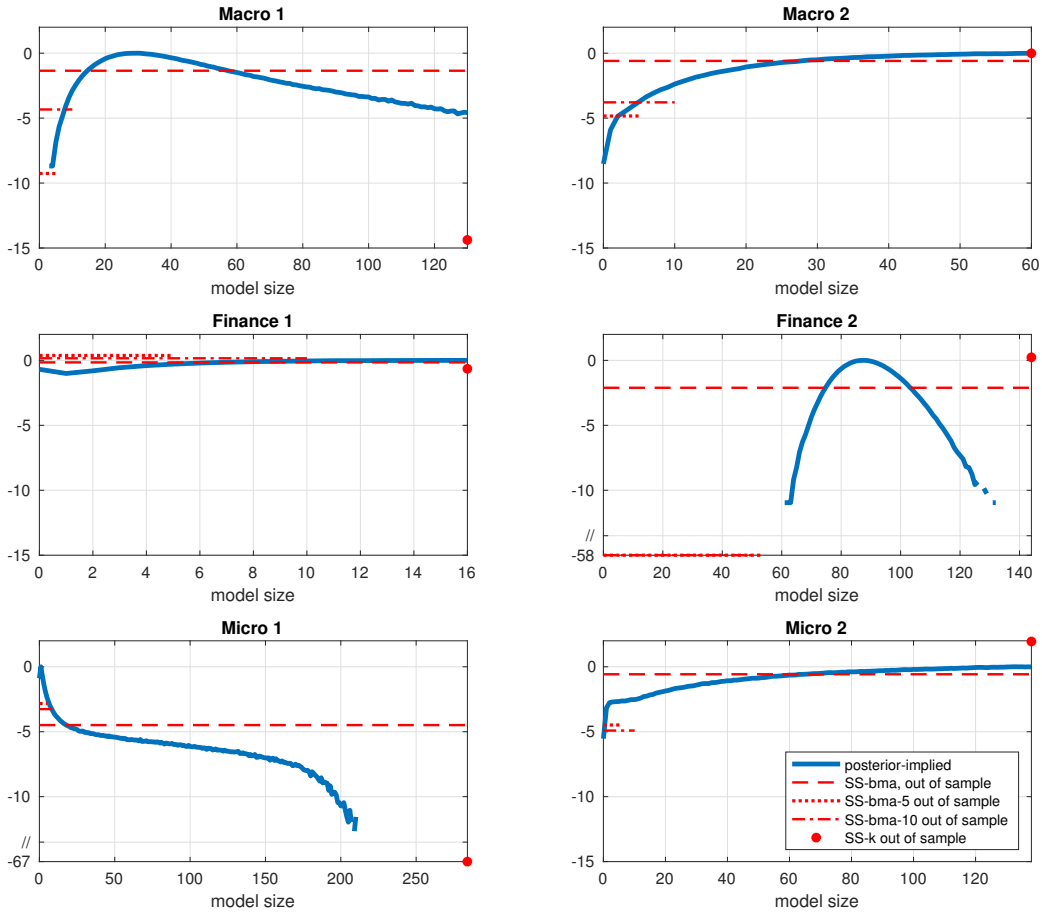


FIGURE 5.1. Log-predictive score relative to the best fitting model. The six panels are plotted using the same y-axis scale for comparability. In finance 2, there are no posterior draws with less than 10 predictors, so the lines corresponding to SS-bma-5 and SS-bma-10 extend to the right till the size of the smallest models visited by the MCMC algorithm.

up to five and ten predictors respectively, weighted by their relative posterior probability (dotted and dashed-dotted lines); and SS-k, which is the dense model with all the predictors, i.e. the ridge regression (the bold dot). For comparability with the posterior of $s$, these new results are also reported in deviation from $\log p(y|s^*)$. What is remarkable is that these "real" out-of-sample forecasting results are very much aligned with those based on the posterior of $s$, not only qualitatively, but also in terms of relative magnitudes. In sum, high-posterior values of $s$ are typically associated with superior out-of-sample predictive

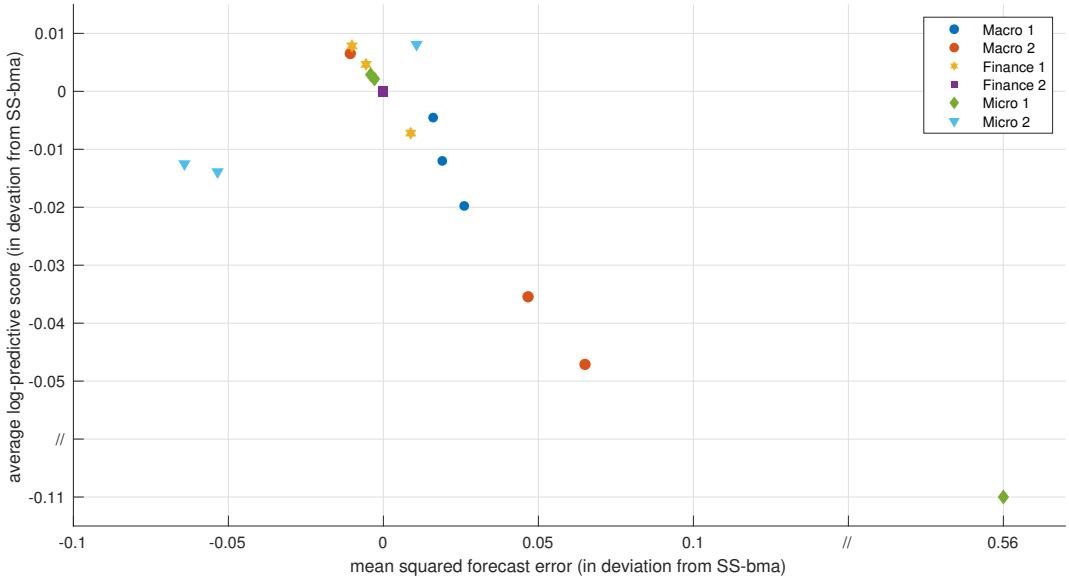


FIGURE 5.2. Relationship between the mean squared forecast errors and the average log-predictive scores of of SS-bma-5, SS-bma-10 and SS-k, in deviations from SS-bma (best viewed in color).

performance, not only according to the theoretical insight of equation (5.1), but also in practice.

Roughly speaking, differences in log-predictive scores across models can be due to differences in point forecasts and/or in the dispersion of the predictive density. To determine whether the accuracy of point and density forecasts are aligned with each other, figure 5.2 plots the mean squared forecast errors of SS-bma-5, SS-bma-10 and SS-k (relative to SS-bma) against their average log-predictive scores (also relative to SS-bma) for our six applications. Observe that there is a clear negative correlation between these two objects, suggesting that models with better density forecasts also have lower mean squared forecast errors. The exception to this rule is micro 2, in which the average log-predictive score of the dense model is high because its density forecasts are more spread out when point forecasts are less accurate.

5.2. **Predictive accuracy and model uncertainty.** In the previous subsection we have documented that small models have lower predictive accuracy, except for the case of micro 1, and that this feature is well captured by the posterior density of $s$. In this subsection we show that the performance of small models deteriorates even further if we ignore model uncertainty and focus on individual low-dimensional models. Figures 5.3 and 5.4 compare

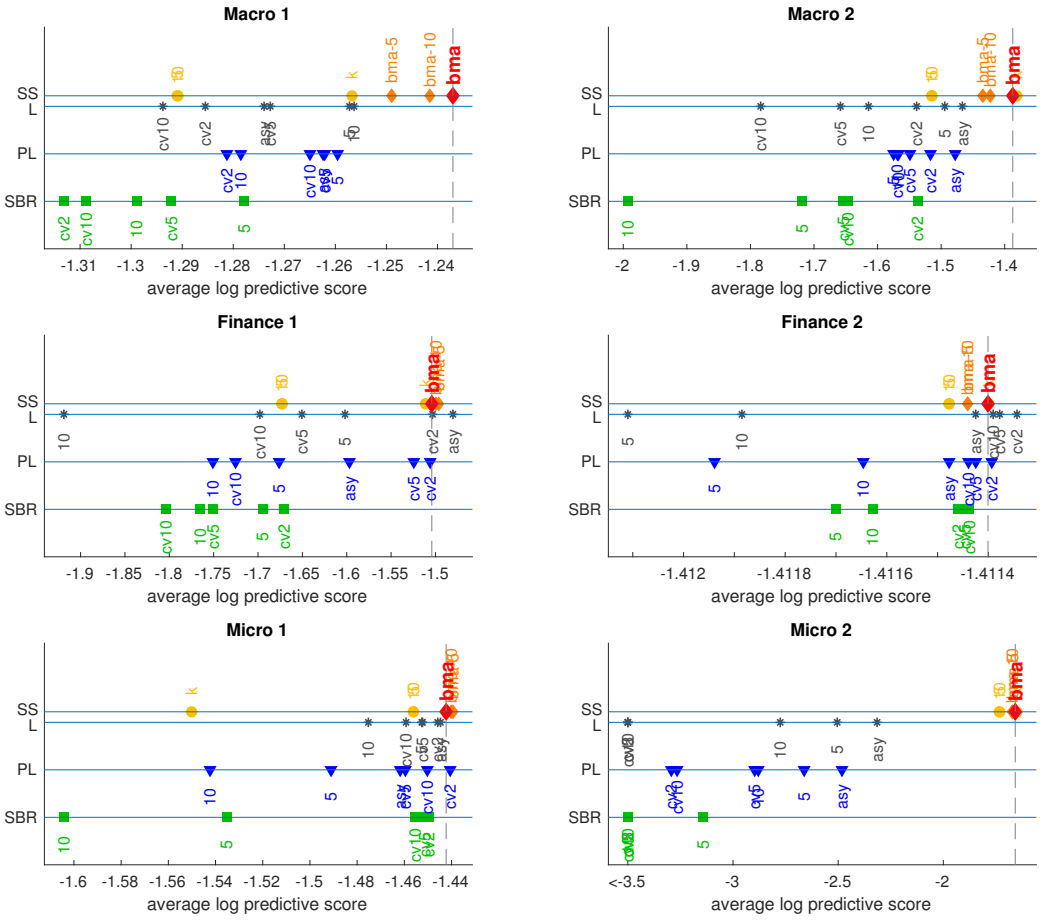


FIGURE 5.3. Out-of-sample average log predictive scores of different models (best viewed in color).

the average log-predictive score and the mean squared forecast error of SS-bma, SS-bma-5 and SS-bma-10 to those of the single best models with up to five and ten predictors. The latter, which we denote by SS-5 and SS-10, are selected as the models with the highest posterior probability in the set of those with up to five and ten predictors.

The upper part of each panel shows that these modal models typically perform worse than the bma strategies. Put differently, in the presence of model uncertainty, a model combination approach is superior to picking individual sparse models, even if these individual models perform well in sample.

The superior predictive performance of model averaging is not an artifact of the way our Bayesian procedure identifies the best individual low-dimensional models, based on

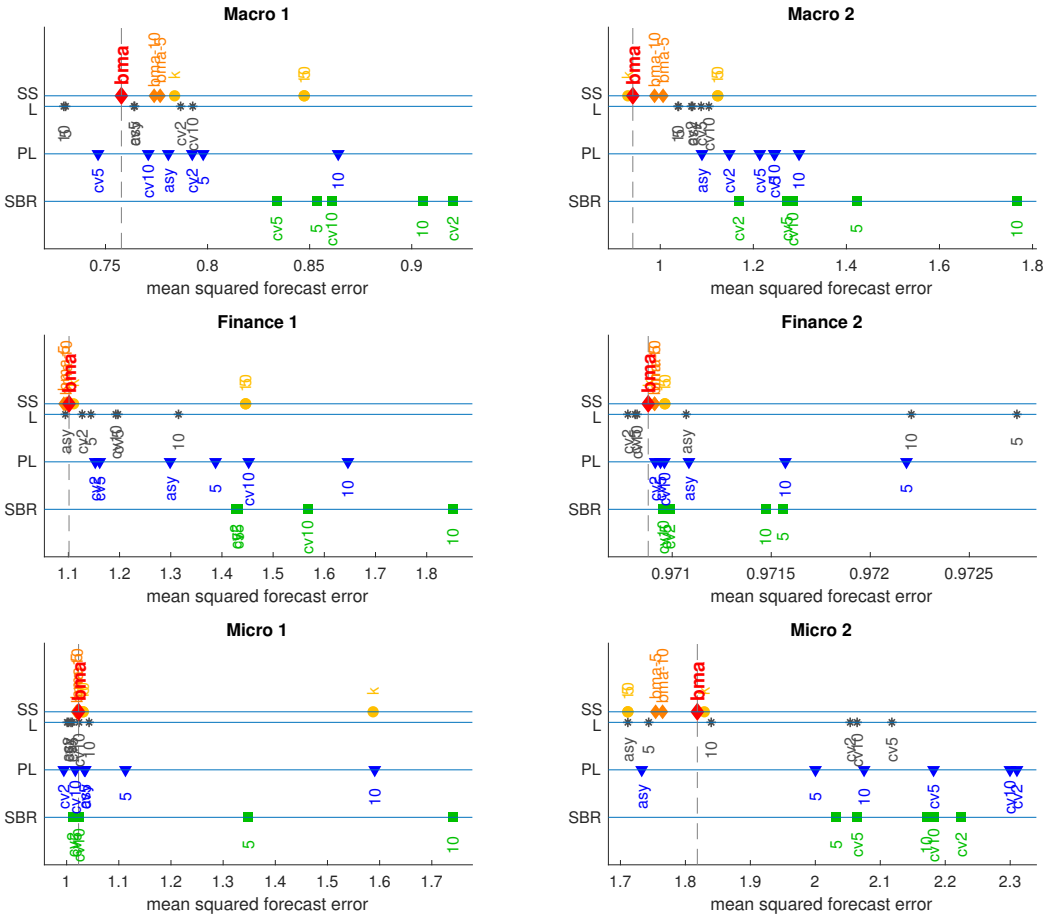


FIGURE 5.4. Out-of-sample mean squared prediction error of different models (best viewed in color).

posterior weights. We substantiate this point in the lower part of each panel, which presents the predictive performance of individual low-dimensional models selected with a variety of alternative methods used in the literature to approximate the solution of regressions with $L_0$ penalty. These methods include many variants of single-best-replacement, lasso and post-lasso techniques, with a fixed number of predictors, and with selection based on asymptotic criteria and cross validation.[14] There are only a few cases in which these approaches marginally improve over the modeling averaging strategy. Therefore, by and large, these results reinforce the point that individual sparse models are typically associated with predictive losses.

[14]See appendix C for references and details.

| | Macro 1 | Macro 2 | Finance 1 | Finance 2 | Micro 1 | Micro 2 |
|---|---|---|---|---|---|---|
| $100 \cdot (\text{ALPS}_{\text{best}} - \text{ALPS}_{\text{SS-bma}})$ | 7.13 | 9.71 | 8.09 | 0.02 | 1.42 | 12.16 |
| $100 \cdot (\text{MSFE}_{\text{SS-bma}}/\text{MSFE}_{\text{best}} - 1)$ | 25.24 | 29.35 | 15.95 | 0.03 | 4.65 | 45.30 |

TABLE 2. Comparison between the predictive accuracy of SS-bma and the ex-post best model. $ALPS$ and $MSFE$ denote the average log predictive score and the mean squared forecast error, respectively.

As a final exercise, table 2 compares the predictive accuracy of the SS-bma strategy to that of the "ex-post" best model in each application. We select such model as the one with the highest out-of-sample average log predictive score or lowest mean squared forecast error among those visited by the MCMC algorithm in the full-sample estimation.[15] Notice that this ex-post best model is an *unfeasible* benchmark, since we can identify it only with the benefit of hindsight. Nevertheless, this comparison is informative because its performance constitutes an upper bound for the out-of-sample predictive accuracy of the forecasting procedures involving *feasible* model selection or model averaging steps. The table shows that the SS-bma strategy is relatively close to this unfeasible upper bound, especially in terms of average log predictive scores measuring the quality of density predictions. However, the gaps in forecasting performance reported in table 2 should be interpreted with caution, because they also measure the extent to which the ex-post best model suffers from overfitting. This is the case because the latter is selected only ex-post, based on its performance on all test samples of our forecasting exercise, and it is not a model that would be chosen in "real time," i.e. given the information in any single training sample. As such, the selection of the ex-post best model is akin to maximizing in-sample—as opposed to out-of-sample—fit over the model space.

## 6. Concluding Remarks

In economics, there is no theoretical argument suggesting that predictive models should in general include only a handful of predictors. As a consequence, the use of low-dimensional model representations can be justified only when supported by strong statistical evidence.

[15] Ideally, we should search over all possible models, but this set is too large for the problem to be manageable. Therefore, we limit the search to models with positive posterior probability, which is likely to include the models with the best predictive accuracy.

In this paper, we evaluate this evidence by studying a variety of predictive problems in macroeconomics, microeconomics and finance. Our main finding is that the empirical support for low-dimensional models is generally weak. Even when it appears stronger, economic data are not informative enough to uniquely identify the relevant predictors when a large pool of variables is available to the researcher. Put differently, predictive model uncertainty seems too pervasive to be treated as statistically negligible. The right approach to scientific reporting is thus to assess and fully convey this uncertainty, rather than understating it through the use of dogmatic (prior) assumptions favoring low dimensional models.

## Appendix A. Algorithm for Posterior Inference

To estimate the model, it is useful to rewrite it using a set of latent variables $z = [z_1, ..., z_k]'$ that are equal to 1 when the corresponding regressor is included in the model and its coefficient is non-zero. Let us define $Y = [y_1, ..., y_T]'$, $U = [u_1, ..., u_T]'$ and $X = [x_1, ..., x_T]'$, where $T$ is the number of observations. The posterior of the unknown objects of the model is given by

$$
\begin{aligned}
p\left(\phi, \beta, \sigma^2, R^2, z, q | Y, U, X\right) &\propto p\left(Y | U, X, \phi, \beta, \sigma^2, R^2, z, q\right) \cdot p\left(\phi, \beta, \sigma^2, R^2, z, q\right) \\
&\propto p\left(Y | U, X, \phi, \beta, \sigma^2\right) \cdot p\left(\beta | \sigma^2, R^2, z, q\right) \cdot p\left(z | q, \sigma^2, R^2\right) \cdot p\left(q\right) \cdot p\left(\sigma^2\right) \cdot p\left(R^2\right) \\
&\propto \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma^2}(Y-U\phi-X\beta)'(Y-U\phi-X\beta)} \\
&\cdot \prod_{i=1}^{k} \left[\left(\frac{1}{2\pi\sigma^2\gamma^2}\right)^{\frac{1}{2}} e^{-\frac{\beta_i^2}{2\sigma^2\gamma^2}}\right]^{z_i} \left[\delta\left(\beta_i\right)\right]^{1-z_i} \\
&\cdot \prod_{i=1}^{k} q^{z_i} \left(1-q\right)^{1-z_i} \\
&\cdot q^{a-1} \left(1-q\right)^{b-1} \\
&\cdot \left(\frac{1}{\sigma^2}\right) \\
&\cdot \left(R^2\right)^{A-1} \left(1-R^2\right)^{B-1},
\end{aligned}
$$

where $\gamma^2 = \frac{1}{k\ \bar{v}_x\ q} \cdot \frac{R^2}{1-R^2}$, and $\delta\left(\cdot\right)$ is the Dirac-delta function.

We can sample from the posterior of $\left(\phi, \beta, \sigma^2, R^2, z, q\right)$ using a Gibbs sampling algorithm with blocks (i) $\left(R^2, q\right)$, (ii) $\phi$, and (iii) $\left(z, \beta, \sigma^2\right)$.

- The conditional posterior of $R^2$ and $q$ is given by

$$
\begin{aligned}
p\left(R^2, q | Y, U, X, \phi, \beta, \sigma^2, z\right) &\propto \left[e^{-\frac{1}{2\sigma^2}\frac{k\ \bar{v}_x\ q\ \left(1-R^2\right)}{R^2}\beta' diag(z)\beta}\right] \cdot \\
&\cdot q^{s(z)+\frac{s(z)}{2}+a-1} \left(1-q\right)^{k-s(z)+b-1} \cdot \left(R^2\right)^{A-1-\frac{s(z)}{2}} \left(1-R^2\right)^{\frac{s(z)}{2}+B-1},
\end{aligned}
$$

  where $s\left(z\right) \equiv \sum_{i=1}^{k} z_i$. We can sample from this distribution by discretizing the $[0, 1]$ support of $R^2$ and $q$. More specifically, for both $R^2$ and $q$ we define a grid with increments of 0.01, and finer increments of 0.001 near the boundaries of the support.

- The conditional posterior of $\phi$ is given by

$$p\left(\phi|Y,U,X,z,\beta,R^2,q,\sigma\right) \propto e^{-\frac{1}{2\sigma^2}(Y-U\phi-X\beta)'(Y-U\phi-X\beta)},$$

which implies

$$\phi|Y,U,X,z,\beta,\gamma,q,\sigma \sim \mathcal{N}\left(\left(U'U\right)^{-1}U'\left(Y-X\beta\right),\sigma^2\left(U'U\right)^{-1}\right).$$

- To draw from the posterior of $z,\beta,\sigma^2|Y,U,X,\phi,R^2,q$, we first draw from $p\left(z|Y,U,X,\phi,R^2,q\right)$, and then from $p\left(\beta,\sigma^2|Y,U,X,\phi,R^2,q,z\right)$. To draw from the posterior of $z|Y,U,X,\phi,R^2,q$, observe that

$$p\left(z|Y,U,X,\phi,R^2,q\right) = \int p\left(z,\beta,\sigma^2|Y,U,X,\phi,R^2,q\right)d\left(\beta,\sigma^2\right)$$

$$\propto q^{s(z)}\left(1-q\right)^{k-s(z)}\left(\frac{1}{2\pi\gamma^2}\right)^{\frac{s(z)}{2}}\int\left(\frac{1}{\sigma^2}\right)^{\frac{T+s(z)}{2}+1}e^{-\frac{1}{2\sigma^2}\left[\left(Y-U\phi-\tilde{X}\tilde{\beta}\right)'\left(Y-U\phi-\tilde{X}\tilde{\beta}\right)+\tilde{\beta}'\tilde{\beta}/\gamma^2\right]}d\left(\tilde{\beta},\sigma^2\right)$$

$$\propto q^{s(z)}\left(1-q\right)^{k-s(z)}\left(\frac{1}{2\pi\gamma^2}\right)^{\frac{s(z)}{2}}\left(2\pi\right)^{\frac{s(z)}{2}}\left|\tilde{W}\right|^{-\frac{1}{2}}\int\left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}+1}e^{-\frac{1}{2\sigma^2}\left[\tilde{Y}'\tilde{Y}-\hat{\tilde{\beta}}'\tilde{W}\hat{\tilde{\beta}}\right]}d\sigma^2$$

$$\propto q^{s(z)}\left(1-q\right)^{k-s(z)}\left(\frac{1}{\gamma^2}\right)^{\frac{s(z)}{2}}\left|\tilde{W}\right|^{-\frac{1}{2}}\left[\frac{\tilde{Y}'\tilde{Y}-\hat{\tilde{\beta}}'\tilde{W}\hat{\tilde{\beta}}}{2}\right]^{-\frac{T}{2}}\Gamma\left(\frac{T}{2}\right),$$

where $\tilde{\beta}$ is the vector of the non-zero coefficients (i.e. those corresponding to $z_i=1$), $\tilde{X}$ are the corresponding regressors, $\hat{\tilde{\beta}}=\tilde{W}^{-1}\tilde{X}'\tilde{Y}$, $\tilde{W}=\left(\tilde{X}'\tilde{X}+I_{\tau(z)}/\gamma^2\right)$, and $\tilde{Y}=Y-U\phi$. Therefore, to draw from the posterior of $z|Y,U,X,\phi,R^2,q$, we can use a Gibbs sampler that allows to draw from the distribution of $z_i|Y,U,X,\phi,R^2,q,z_{-i}$. Finally, to draw from the posterior of $\beta,\sigma^2|Y,U,X,\phi,R^2,q,z$, observe that

$$\sigma^2|Y,U,X,\phi,R^2,q,z \sim IG\left(\frac{T}{2},\frac{\tilde{Y}'\tilde{Y}-\hat{\tilde{\beta}}'\left(\tilde{X}'\tilde{X}+I_{s(z)}/\gamma^2\right)\hat{\tilde{\beta}}}{2}\right)$$

and

$$\tilde{\beta}|Y,U,X,\phi,\sigma^2,R^2,q,z \sim \mathcal{N}\left(\hat{\tilde{\beta}},\sigma^2\left(\tilde{X}'\tilde{X}+I_{s(z)}/\gamma^2\right)^{-1}\right),$$

and the other $\beta_i$'s are equal to 0.

## Appendix B. Additional Simulations

This appendix expands the simulation evidence of section 2.1, by considering alternative designs in which the regression coefficients are drawn from a Laplace distribution or from mixtures of Gaussian distributions with a bimodal shape. These simulations are otherwise identical to the second set of simulations described in section 2.1 and figure 2.2, i.e. they include non-Gaussian and heteroskedastic disturbances.

Figure B.1 considers the case in which the non-zero regression coefficients are drawn from a Laplace distribution with mean zero and variance equal to one. Relative to a Gaussian, the Laplace density has more mass around zero and in the tails. Figure B.2 analyzes instead the outcome of simulations with non-zero coefficients drawn from a mixture of two Gaussian distributions. The first component of the mixture is a Gaussian with mean equal to $-2/\sqrt{5}$ and variance $1/5$. The second mixture component is equal to the first, but its mean is $2/\sqrt{5}$. The mixture weights are equal to $1/2$. The resulting mixture distribution has mean zero and variance equal to one, and it is bimodal. Finally, figure B.3 studies the case in which the non-zero regression coefficients are drawn from a mixture of Gaussian distributions with positive mean. This mixture is similar to the one just described, except for the fact that the means of the two components are 0 and $4/\sqrt{5}$, so that the overall mean and variance of the distribution are $2/\sqrt{5}$ and 1. Figures B.1, B.2 and B.3 show that the model continues to detect the true level of sparsity quite well, and its performance is thus not particularly sensitive to the exact distribution of the non-zero regression coefficients.

## Appendix C. Details of the Out-of-Sample Prediction Exercise

The out-of-sample prediction exercise is designed as a standard forecasting exercise for applications with time-series data, as a cross-validation exercise for applications with cross-sectional data, and a combination of the two for applications with panel data. The details are as follows:

- **Macro 1**. We estimate the model on data from 1960:2 to 1974:12, evaluating its one-month-ahead forecasting accuracy over the subsequent year of the sample, from 1975:1 to 1975:12. We repeat this exercise 44 times, by adding each time one year of data to the training sample and shifting the evaluation sample by one year.[16]

[16]The time series dimension of the smallest training sample is approximately 25 percent of the full time series dimension. We followed this approximate rule in all applications.

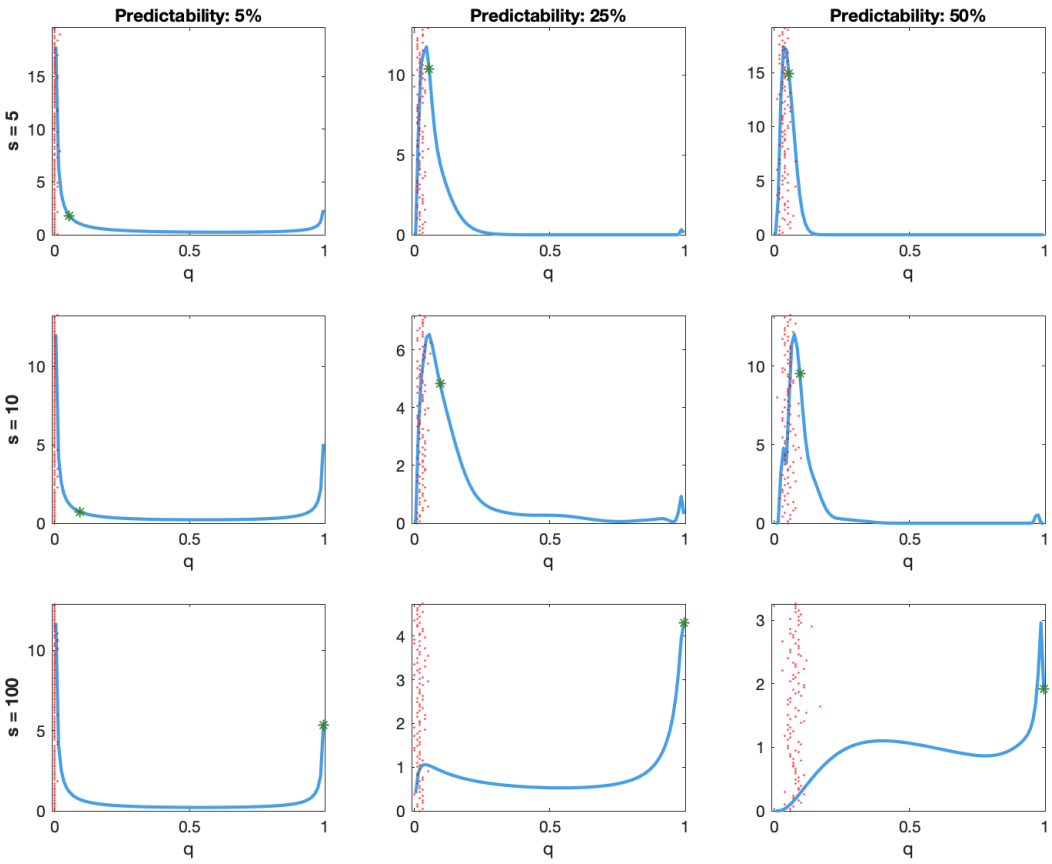


FIGURE B.1. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a Laplace distribution: Kernel approximation of the distribution of the posterior mode of $q$ across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by a lasso regression, with penalty parameter based on the asymptotically optimal criterion proposed by Bickel et al. (2009) and the tuning constants recommended by Belloni et al. (2011a) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

- **Macro 2**. We estimate the model on a randomly selected sample of 50 percent of the countries, evaluating prediction accuracy on the remaining 50 percent of the observations. We repeat this exercise 100 times.
- **Finance 1**. We estimate the model on data from 1948 to 1964, evaluating the accuracy of the forecast of the 1965 observation. We repeat this exercise 51 times, by adding each time one yearly observation to the training sample and shifting the evaluation sample by one year.

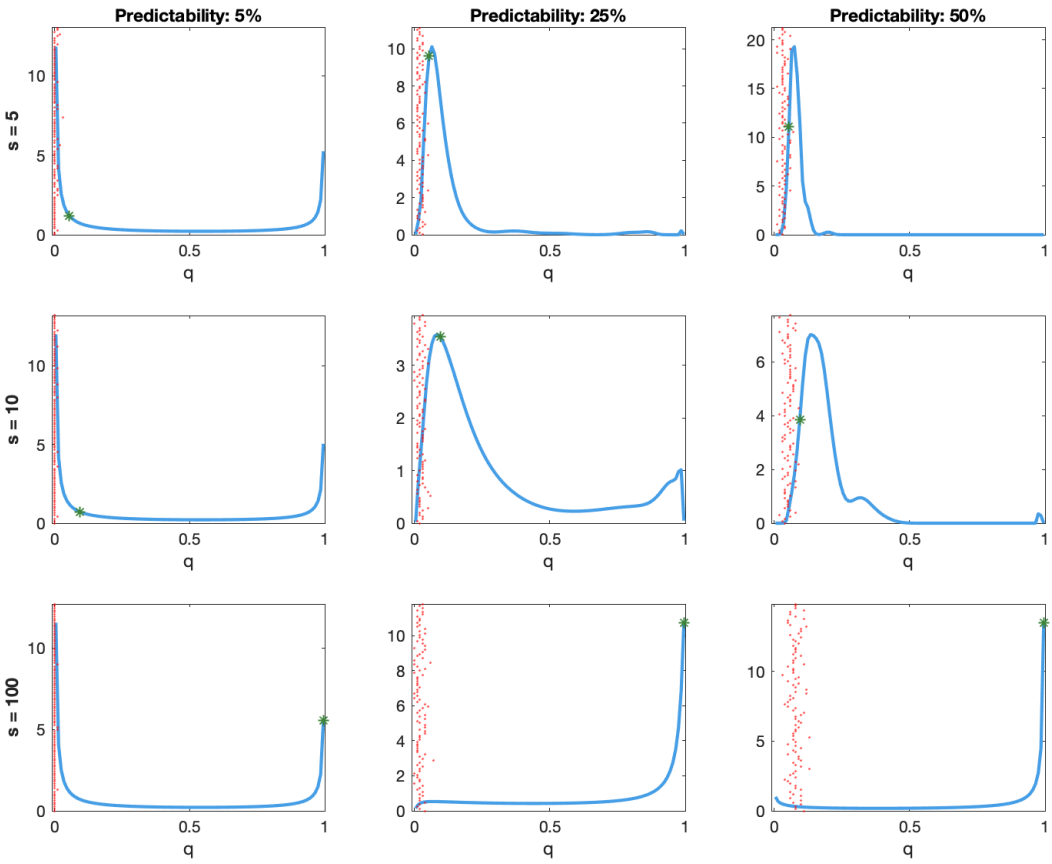


FIGURE B.2. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a zero-mean mixture of Gaussians: Kernel approximation of the distribution of the posterior mode of $q$ across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by a lasso regression, with penalty parameter based on the asymptotically optimal criterion proposed by Bickel et al. (2009) and the tuning constants recommended by Belloni et al. (2011a) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

- **Finance 2**. We estimate the model on data from 1963:1 to 1974:12, evaluating the one-month-ahead forecasting accuracy of all the stock returns over the subsequent year, from 1975:1 to 1975:12. We repeat this exercise 40 times, by adding each time one year of data to the training sample and shifting the evaluation sample by one year.
- **Micro 1**. We estimate the model using all the data for a randomly selected sample of 50 percent of the states (group 1), and data from 1986 to 1989 for the remaining

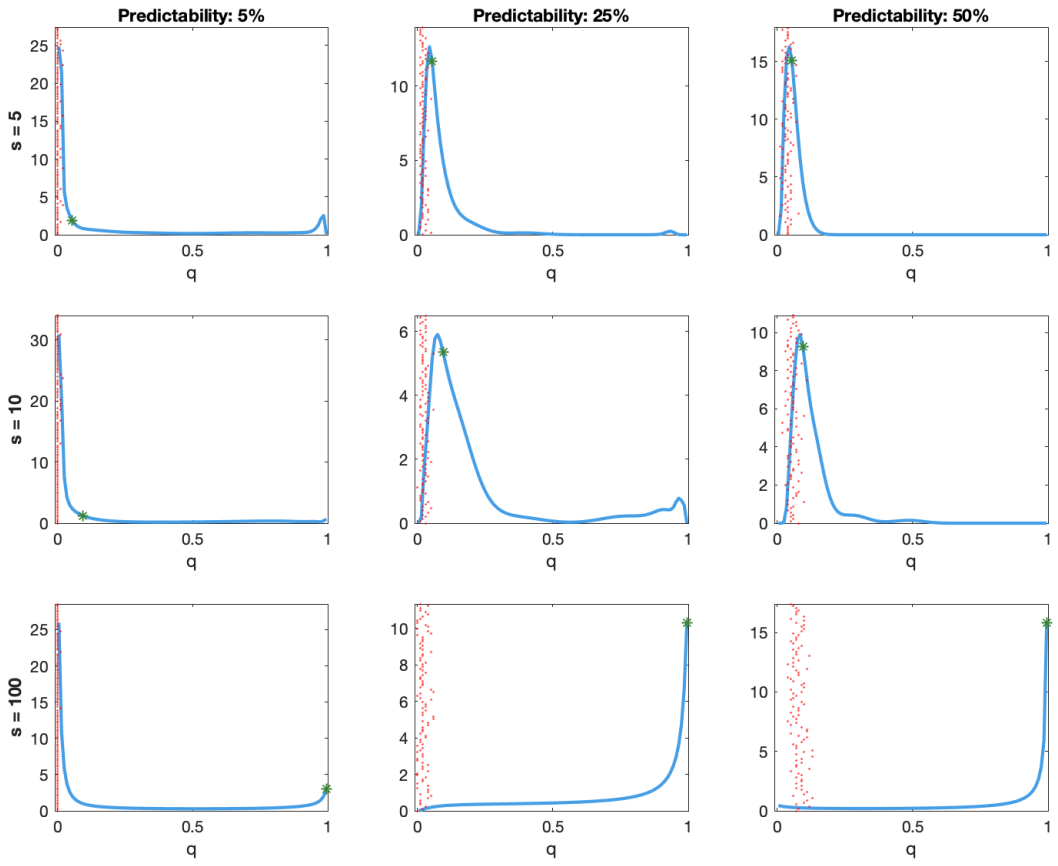


FIGURE B.3. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a positive-mean mixture of Gaussians: Kernel approximation of the distribution of the posterior mode of $q$ across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by a lasso regression, with penalty parameter based on the asymptotically optimal criterion proposed by Bickel et al. (2009) and the tuning constants recommended by Belloni et al. (2011a) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

50 percent of the states (group 2). This "mixed" strategy to form a training sample is necessary because the model involves a full set of year dummies. We evaluate the accuracy of the model predictions for the second group of states in year 1990. We repeat this procedure 8 times (including the random split of the states into group 1 and 2), by adding each time one year of data for group 2 and shifting the evaluation sample by one year. Finally, we repeat the whole exercise 13 times, for a total of 104 training and test samples.

- **Micro 2**. We estimate the model using all the data for a randomly selected sample of 50 percent of the circuits (group 1), and data from 1979 to 1984 for the remaining 50 percent of the circuits (group 2). This "mixed" strategy to form a training sample is necessary because the model involves a full set of year and circuit dummies. We evaluate the accuracy of the model predictions for the second group of circuits in year 1985. We repeat this procedure 20 times (including the random split of the circuits into group 1 and 2), by adding each time one year of data for group 2 and shifting the evaluation sample by one year. Finally, we repeat the whole exercise 5 times, for a total of 100 training and test samples.

The measures of forecasting accuracy reported in the main text are computed by averaging the log-predictive scores and the squared forecast errors over the elements of a test sample, and across all test samples.

We evaluate the prediction accuracy of the following baseline and restricted versions of our model: SS-bma, which is our full model that combines all the possible individual models, weighted by their posterior probability; SS-bma-5 and SS-bma-10, which restrict the model space to the combinations of individual models with up to five and ten predictors respectively, weighted by their relative posterior probability; and SS-k, which is the dense model including all the predictors. The predictive density of $y_{T+1}$ implied by these models is a mixture of Gaussian densities with means $u'_{T+1}\phi^{(j)} + x'_{T+1}\beta^{(j)}$ and variances $\sigma^{2(j)}$, where $\phi^{(j)}$, $\beta^{(j)}$ and $\sigma^{2(j)}$, $j = 1, ..., M$, are draws from their posterior distribution. The predictive score is computed as the value of this density at the actual realization of $y_{T+1}$. We use the mean of the predictive density as the point forecast for the computation of the mean squared forecast error (with the exception of micro 2, for which we use the mode of the density evaluated at the three possible values of the response variable in this application).

To select the "best" individual models for each training sample, we employ three different sparse modeling strategies:

- **Spike-and-slab (SS)**. Within our spike-and-slab framework, we select SS-5 and SS-10 as the individual models with the highest posterior probability in the set of those with up to five and ten predictors. To robustify the procedure, instead of simply counting the number of times an individual model is visited by the MCMC algorithm, we numerically compute the posterior model probability of all models

that are visited at least once, and pick the model with the highest.[17] The predictive density of $y_{T+1}$ implied by these models is a mixture of Gaussian densities with means $u'_{T+1}\phi^{(j)} + x'_{T+1}\beta^{(j)}$ and variances $\sigma^{2(j)}$, where $\phi^{(j)}$, $\beta^{(j)}$ and $\sigma^{2(j)}$, $j = 1, ..., M$, are draws from their posterior distribution. We use the mean of the predictive density as the point forecast for the computation of the mean squared forecast error.

- **Lasso (L)** and **Post-lasso (PL)**. As an alternative way to identify good-fitting individual small models, we also consider the popular lasso method (Tibshirani, 1996). We consider the following variants of this methodology. (i) L-5 and L-10: lasso with a fixed number of five and ten predictors; (ii) L-asy: lasso with a penalty parameter based on the asymptotic criterion proposed by Bickel et al. (2009), implemented using the iterative procedure and the tuning constants recommended by Belloni et al. (2011a) (notice that this criterion is designed for valid inference, not necessarily best prediction); (iii) L-cv2, L-cv5 and L-cv10: lasso with selection of the number of predictors based on 2-, 5- and 10-fold cross validation.[18] It is well known that constructing the full predictive density implied by lasso is challenging, and there is no agreement in the literature about how to tackle this problem (Hastie et al., 2015). For this reason, we use two alternative rough approximations of the density of $y_{T+1}$.

  The first method consists of treating the lasso parameter estimates as known, and assuming Gaussian errors and a flat prior on their variance. Under these assumptions, the density of $y_{T+1}$ is a non-centered Student-$t$ distribution, with mean $u'_{T+1}\hat{\phi}_L + x'_{T+1}\hat{\beta}_L$, scale $\sqrt{\hat{r}_L/(T-2)}$ and degrees of freedom $T-2$, where $\hat{\phi}_L$, $\hat{\beta}_L$ and $\hat{r}_L$ are the lasso estimates of $\phi$, $\beta$ and the sum of squared residuals. As before, we use the mean of the predictive density ($u'_{T+1}\hat{\phi}_L + x'_{T+1}\hat{\beta}_L$) as the point forecast for the computation of the mean squared forecast error (with the exception of micro 2, for which we use the mode of the density evaluated at the three possible values of the response variable in this application).

[17] If models with less than 5 or 10 predictors receive less than 0.05 percent of the total posterior weight, we consider progressively larger models until we reach this lower bound. The only application where this is an issue is finance 2, where small models are essentially never visited.

[18] We approximate the lasso estimates with five and ten predictors with the fifth and tenth step of the least-angle regression (LARS) algorithm. Similarly, for the case of cross-validation, we search over the possible number of steps in the LARS algorithm as opposed to the values of the penalty, to improve speed.

An alternative method to construct the predictive density is based on post-selection inference. It consists of running a simple ordinary least squares regression of the response variable on the regressors selected by lasso (Belloni and Chernozhukov, 2013). This "post-lasso" procedure reduces the bias of the lasso estimator and may better approximate the solution of the best subset selection problem (Beale et al., 1967 and Hocking and Leslie, 1967). With Gaussian errors and a flat prior on the second-stage regression, the implied predictive density of $y_{T+1}$ is a non-centered Student-$t$ distribution, with mean $u'_{T+1}\hat{\phi}_{PL} + x'_{T+1}\hat{\beta}_{PL}$, scale $\sqrt{\left([u'_{T+1}, x'_{T+1}]\left([U,X]'[U,X]\right)^{-1}[u'_{T+1}, x'_{T+1}]' + 1\right)\hat{r}_{PL}/(T-l-n-2)}$ and degrees of freedom $T-l-n-2$, where $\hat{\phi}_{PL}$, $\hat{\beta}_{PL}$ and $\hat{r}_{PL}$ are the ordinary least squares estimates of $\phi$, $\beta$ and the sum of squared residuals in the second-stage regression, and $n$ is the dimension of the vector $\hat{\beta}_{PL}$. This post-selection approach allows us to incorporate parameter uncertainty in the predictive density, although the parameter estimates in the second stage are of course different from the lasso estimates. It is important to stress that this strategy is appropriate only under the stringent assumptions guaranteeing that model selection does not impact the asymptotic distribution of the parameters estimated in the post-selection step (Bhulmann and van de Geer, 2011; see also Leeb and Potscher, 2005, 2008a,b for a thorough discussion of the fragility of this approach, and Chernozhukov et al., 2015 for a comprehensive review of these topics). In the figures of the paper, we denote the log-predictive scores implied by this method as PL-5, PL-10, PL-asy, PL-cv2, PL-cv5 and PL-cv10, depending on the lasso variant used in the selection stage. For completeness, we also report the mean squared forecast error based on post-lasso, using the mean of the predictive density ($u'_{T+1}\hat{\phi}_{PL} + x'_{T+1}\hat{\beta}_{PL}$) as the point forecast (with the usual exception of micro 2).

- **Single best replacement (SBR)**. This class of methods (also known as forward stepwise) is a fast and scalable approximation of the solution of the best subset selection problem, and thus provides yet another way to choose good-fitting sparse individual models. We use the SBR computation algorithm of Soussen et al. (2011) and Polson and Sun (2019), and consider the following variants of this method. (i) SBR-5 and SBR-10: SBR with a fixed number of five and ten predictors; (ii) SBR-cv2, SBR-cv5 and SBR-cv10: SBR with selection of the number of predictors based

on 2-, 5- and 10-fold cross validation. The predictive density and point forecast of $y_{T+1}$ implied by these models are constructed as in the post-lasso case.

## References


Abadie, A. and M. Kasy (2019): "Choosing among Regularized Estimators in Empirical Economics: The Risk of Machine Learning," *The Review of Economics and Statistics*, 101, 743–762.

Athey, S., M. Bayati, G. Imbens, and Z. Qu (2019): "Ensemble Methods for Causal Effects in Panel Data Settings," *AEA Papers and Proceedings*, 109, 65–70.

Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Financial Economics*, 64, 423 – 458.

Bai, J. and S. Ng (2009): "Boosting Diffusion Indices," *Journal of Applied Econometrics*, 24, 607–629.

Bańbura, M., D. Giannone, and M. Lenza (2015): "Conditional forecasts and scenario analysis with vector autoregressions for large cross-sections," *International Journal of Forecasting*, 31, 739–756.

Barro, R. J. (1991): "Economic Growth in a Cross Section of Countries," *The Quarterly Journal of Economics*, 106, 407–443.

Barro, R. J. and J.-W. Lee (1994): "Sources of economic growth," *Carnegie-Rochester Conference Series on Public Policy*, 40, 1 – 46.

Beale, E. M. L., M. G. Kendall, and D. W. Mann (1967): "The Discarding of Variables in Multivariate Analysis," *Biometrika*, 54, 357–366.

Belloni, A., D. L. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

Belloni, A. and V. Chernozhukov (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19, 521–547.

Belloni, A., V. Chernozhukov, and C. Hansen (2011a): "Inference for high-dimensional sparse econometric models," in *Advances in Economics and Econometrics – World Congress of Econometric Society 2010*.

——— (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608.

Belloni, A., V. Chernozhukov, and L. Wang (2011b): "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, 98, 791–806.

Bhulmann, P. and S. van de Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Publishing Company, Incorporated, 1st ed.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," 37, 1705–1732.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010): "The horseshoe estimator for sparse signals," *Biometrika*, 97, 465–480.

Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015): "Bayesian linear regression with sparse priors," *Annals of Statistics*, 43, 1986–2018.

Chen, D. L. and S. Yeh (2012): "Growth under the shadow of expropriation? The economic impacts of eminent domain," Mimeo, Toulouse School of Economics.

Chernozhukov, V., C. Hansen, and Y. Liao (2017): "A lava attack on the recovery of sums of dense and sparse signals," *The Annals of Statistics*, 45, 39–76.

Chernozhukov, V., C. Hansen, and M. Spindler (2015): "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688.

Chudik, A., G. Kapetanios, and M. H. Pesaran (2018): "A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models," *Econometrica*, 86, 1479–1512.

Cremers, K. M. (2002): "Stock return predictability: A Bayesian model selection perspective," *Review of Financial Studies*, 15, 1223–1249.

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.

Donohue, J. J. and S. D. Levitt (2001): "The impact of legalized abortion on crime," *The Quarterly Journal of Economics*, 116, 379–420.

Draper, N. R. and H. Smith (1966): *Applied Regression Analysis*, John Wiley & Sons.

Faust, J., S. Gilchrist, J. H. Wright, and E. Zakrajsek (2013): "Credit Spreads as Predictors of Real-Time Economic Activity: A Bayesian Model-Averaging Approach," *The Review of Economics and Statistics*, 95, 1501–1519.

Fernandez, C., E. Ley, and M. F. J. Steel (2001): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563–576.

Freyberger, J., A. Neuhierl, and M. Weber (2017): "Dissecting Characteristics Nonparametrically," Working Paper 23227, National Bureau of Economic Research.

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

George, E. I. and R. E. McCulloch (1993): "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Harvey, C. R., Y. Liu, and H. Zhu (2016): "... and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29, 5.

Hastie, T., R. Tibshirani, and J. Friedman (2001): *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY, USA: Springer New York Inc.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical learning with sparsity*, CRC press.

Hocking, R. R. and R. N. Leslie (1967): "Selection of the Best Subset in Regression Analysis," *Technometrics*, 9, 531–540.

Hoerl, A. E. and R. W. Kennard (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Inoue, A. and L. Kilian (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.

Ishwaran, H. and J. S. Rao (2005): "Spike and slab variable selection: Frequentist and Bayesian strategies," *Annals of Statistics*, 33, 730–773.

Jin, S., L. Su, and A. Ullah (2014): "Robustify Financial Time Series Forecasting with Bagging," *Econometric Reviews*, 33, 575–605.

Kozak, S., S. Nagel, and S. Santosh (2017): "Shrinking the Cross Section," NBER Working Papers 24070, National Bureau of Economic Research, Inc.

Lawley, D. N. and A. E. Maxwell (1963): *Factor analysis as a statistical method*, Butterworths London.

Leamer, E. E. (1973): "Multicollinearity: A Bayesian Interpretation," *The Review of Economics and Statistics*, 55, 371–380.

Leeb, H. and B. M. Potscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2008a): "Can One Estimate The Unconditional Distribution Of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338–376.

——— (2008b): "Sparse estimators and the oracle property, or the return of Hodges' estimator," *Journal of Econometrics*, 142, 201–211.

Li, Q. and N. Lin (2010): "The Bayesian elastic net," *Bayesian Analysis*, 5, 151–170.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008): "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

McCracken, M. W. and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34, 574–589.

Mitchell, T. J. and J. J. Beauchamp (1988): "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Ng, S. (2013): *Variable Selection in Predictive Regressions*, Elsevier, vol. 2 of *Handbook of Economic Forecasting*, 752–789.

——— (2014): "Viewpoint: Boosting Recessions," *The Canadian Journal of Economics / Revue canadienne d'Economique*, 47, 1–34.

Park, T. and G. Casella (2008): "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

Pearson, K. (1901): "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, 2, 559–572.

Polson, N. G. and L. Sun (2019): "Bayesian l0-regularized least squares," *Applied Stochastic Models in Business and Industry*, 35.

Rapach, D. and J. Strauss (2010): "Bagging or Combining (or Both)? An Analysis Based on Forecasting U.S. Employment Growth," *Econometric Reviews*, 29, 511–533.

Sala-I-Martin, X., G. Doppelhofer, and R. I. Miller (2004): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *American Economic Review*, 94, 813–835.

Soussen, C., J. Idier, D. Brie, and J. Duan (2011): "From BernoulliÐGaussian Deconvolution to Sparse Signal Restoration," *IEEE Transactions on Signal Processing*, 59, 4572–4584.

Spearman, C. (1904): ""General Intelligence," Objectively Determined and Measured," *The American Journal of Psychology*, 15, 201–292.

Stock, J. H. and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 147–162.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economics Statistics*, 20, 147–162.

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

TIKHONOV, A. N. (1963): "Solution of Incorrectly Formulated Problems and the Regularization Method," *Soviet Mathematics Doklady*, 5, 1035/1038.

VARIAN, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, 3–28.

WAGER, S. AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

WELCH, I. AND A. GOYAL (2008): "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21, 1455–1508.

WRIGHT, J. H. (2009): "Forecasting US inflation by Bayesian model averaging," *Journal of Forecasting*, 28, 131–144.

AMAZON AND CEPR

*E-mail address*: dgiannon2@gmail.com

EUROPEAN CENTRAL BANK AND ECARES

*E-mail address*: michele.lenza@ecb.europa.eu

NORTHWESTERN UNIVERSITY, CEPR AND NBER

*E-mail address*: g-primiceri@northwestern.edu