SUPPLEMENT TO

ECONOMIC PREDICTIONS WITH BIG DATA: THE ILLUSION OF SPARSITY

DOMENICO GIANNONE, MICHELE LENZA, AND GIORGIO E. PRIMICERI

This document contains some additional results and technical details not included in the main body of the paper. In particular, we present: i) more Monte Carlo simulation evidence; ii) the details of our out-of-sample forecasting exercise. This supplement is not self-contained, so readers are advised to read the main paper first.

Appendix A. Additional Simulations

This appendix expands the simulation evidence of section 2.1 of the main paper, 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. In the last part of this appendix, we also study what type of extreme assumptions might alter the overall good performance of the model.

A.1. **Alternative distributions of the nonzero regression coefficients.** Figure A.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 A.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 A.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

Date: January 2021.

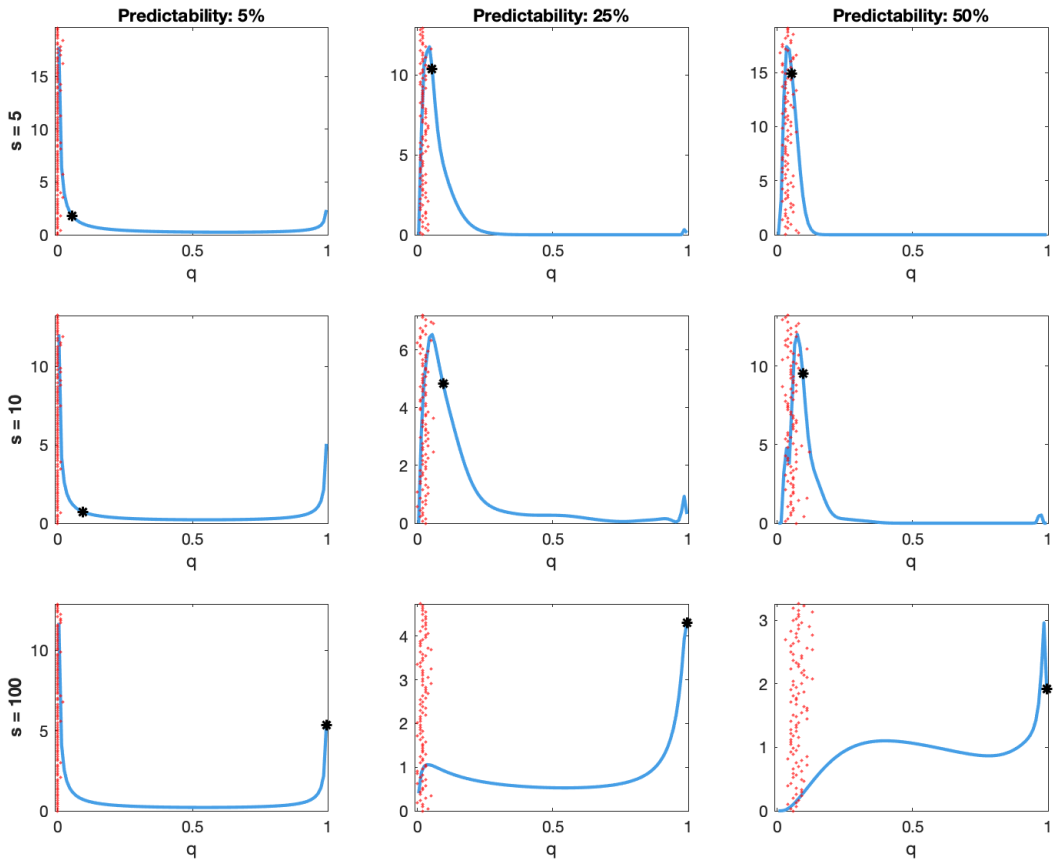

FIGURE A.1. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a Laplace distribution: Kernel approximation of the distribution of the posterior mode of q across simulations (solid line); fraction of non-zero coefficients estimated in each simulation by 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. (2011) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

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 A.1, A.2 and A.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.

A.2. **Testing the boundaries of our model.** These results—in combination with those of section 2.1 in the main paper—are comforting, as they show that our model is able to detect the true level of sparsity even if its parametric assumptions are substantially

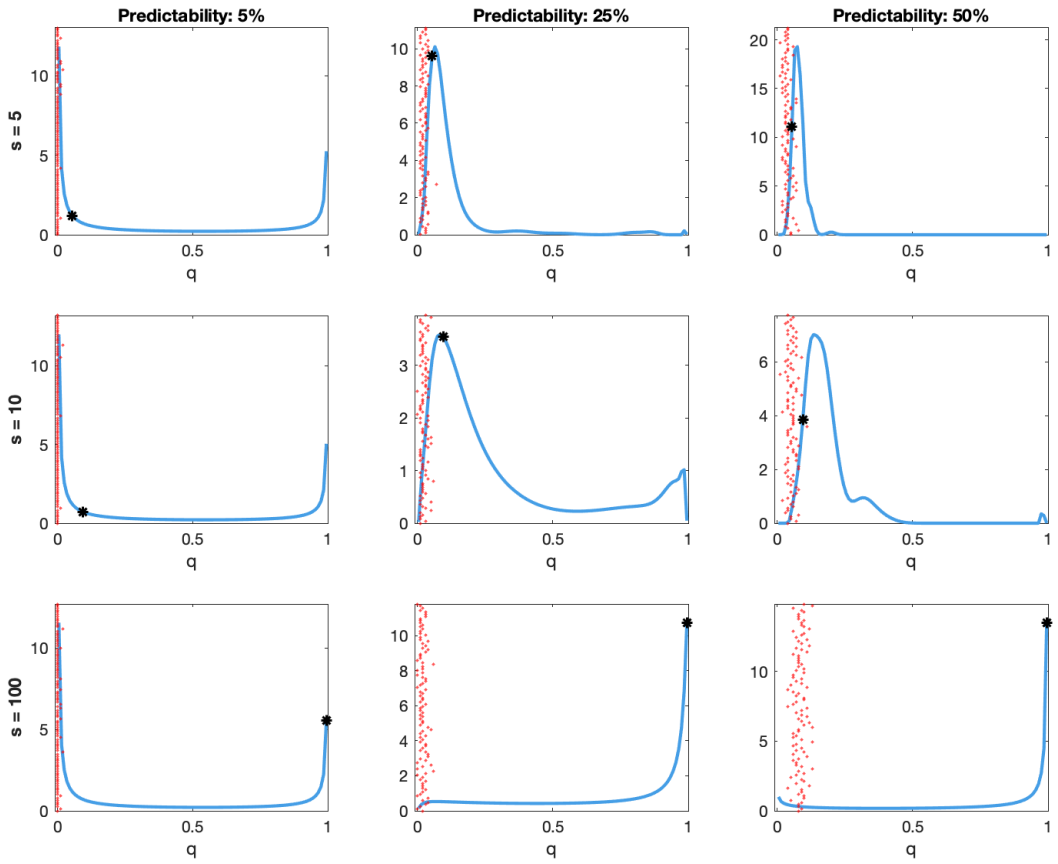

FIGURE A.2. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a zero-mean mixture of Gaussians: Kernel approximation of the distribution of the posterior mode of q across simulations (solid line); fraction of non-zero coefficients estimated in each simulation 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. (2011) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

different from those of the DGP. What type of extreme assumptions about the DGP could then undermine the 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. (2011). Specifically,

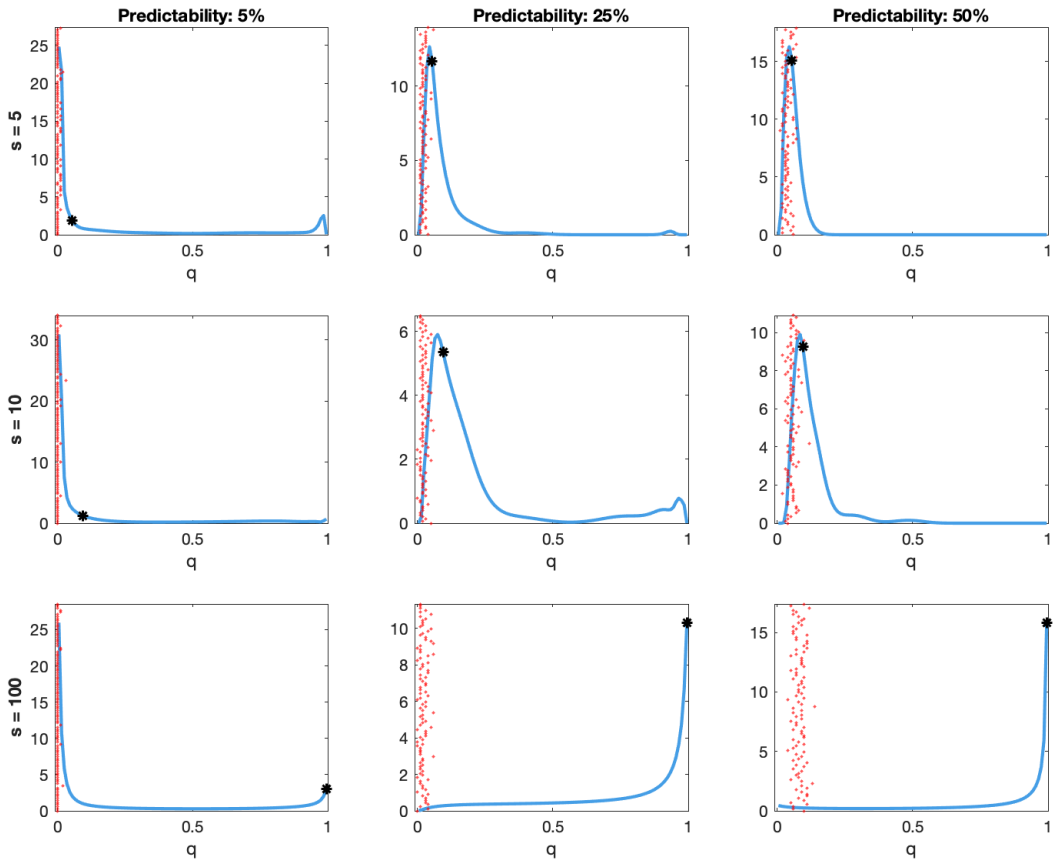

FIGURE A.3. Simulations with non-Gaussian and heteroskedastic errors, and with non-zero coefficients drawn from a positive-mean mixture of Gaussians: Kernel approximation of the distribution of the posterior mode of q across simulations (solid line); fraction of non-zero coefficients estimated in each 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. (2011) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

we repeat the baseline simulations of figure 2.1 with $s = 5$. However, instead of setting 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 A.4, the model is still able to detect the true level of sparsity quite well, even though sparsity is now contaminated by noise.

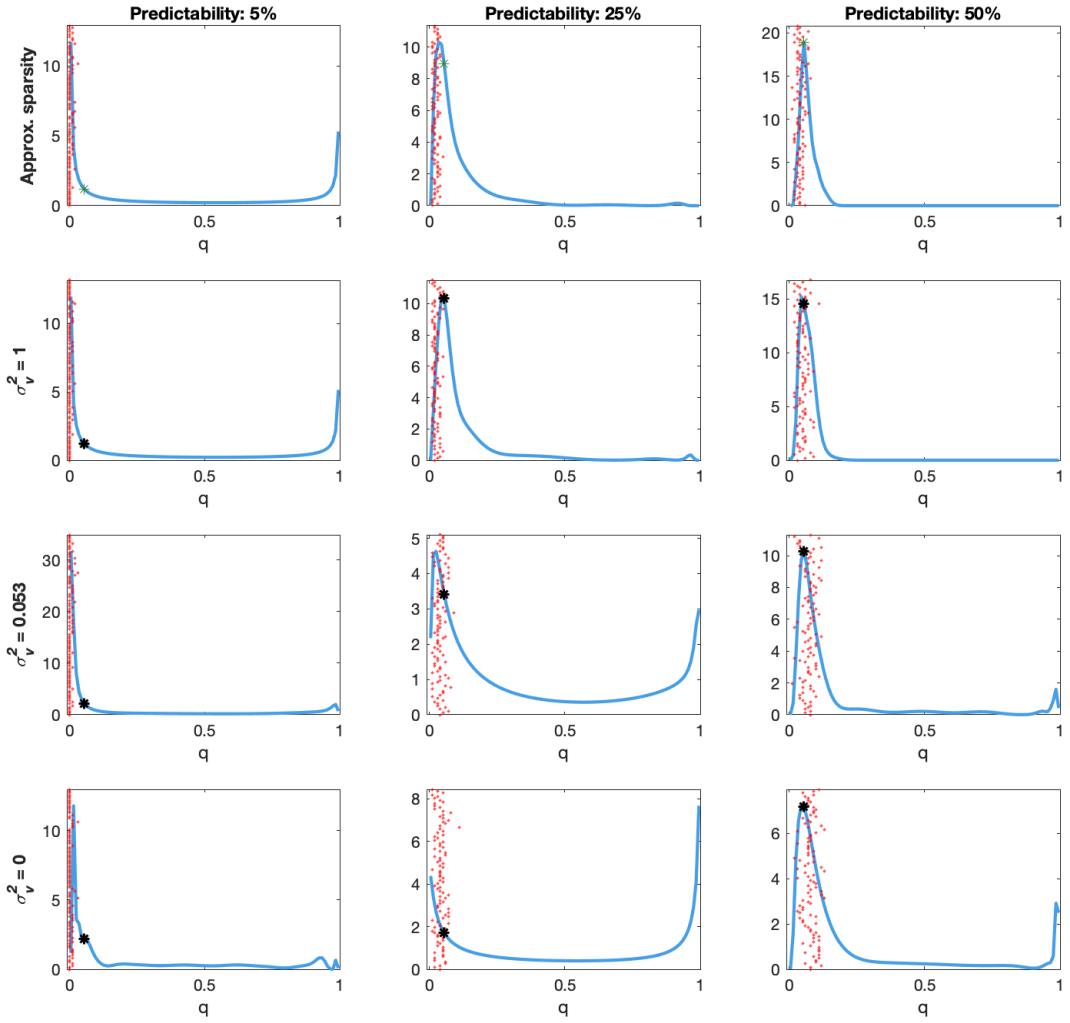

FIGURE A.4. Simulations with approximate sparsity (first row) and a factor structure for the predictors (second, third and fourth rows): Kernel approximation of the distribution of the posterior mode of q across simulations (solid line); fraction of non-zero coefficients estimated in each simulation 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. (2011) (dots); and fraction of non-zero coefficients in each simulation design (starred dot).

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.[1] 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 Λ 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 σ_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 A.4 present the outcome of this last set of simulations. When σ_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 σ_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 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

[1] 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.

of sparsity, given that these last simulations are exactly designed so that a dense model is a good approximation of the sparse one.

Appendix B. Details of the Out-of-Sample Prediction Exercise

This appendix provides the details of the out-of-sample exercise presented in the main text. This 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 as a combination of the two for the micro 1 and micro 2 applications.[2] Table 1—which is also reported in the main body of the paper—provides some details about the construction of the training and evaluation samples in each of the six applications.

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: BMA-all, which is our full model that combines all the possible individual models, weighted by their posterior probability; BMA-5 and 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

[2]Combining time series and cross-section cross-validation strategies is necessary in the case of micro 1 and micro 2, because these applications involve full sets of year, and year and circuit dummies, respectively.

	Training and evaluation samples
Macro 1	We estimate the model on data from 1960:2 to 1969:12, evaluating its predictions for the 1970:1–1970:12 observations. We repeat this exercise 45 times, each time expanding the training sample by one year and shifting the evaluation sample accordingly.
Macro 2	We estimate the model on a randomly selected sample of 50 percent of the countries, evaluating its predictions for 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 its prediction for the 1965 observation. We repeat this exercise 51 times, each time expanding the training sample by one year and shifting the evaluation sample accordingly.
Finance 2	We estimate the model on data from 1963:1 to 1974:12, evaluating its predictions for the 1975:1–1975:12 observations of all the stock returns. We repeat this exercise 40 times, each time expanding the training sample by one year and shifting the evaluation sample accordingly.
Micro 1	We estimate the model using data from 1986 to 1989 for all states, and from 1990 to 1997 for a 50 percent randomly selected sample of states. We evaluate the model predictions for this last group of states in year 1990. We repeat this procedure 8 times (including the random selection of 50 percent of the states), each time adding one year of data to the training sample and shifting the evaluation sample accordingly. We repeat the whole exercise 13 times, for a total of 104 training and evaluation samples.
Micro 2	We estimate the model using data from 1979 to 1984 for all circuits, and from 1985 to 2004 for a 50 percent randomly selected sample of circuits. We evaluate the model predictions for this last group of circuits in year 1985. We repeat this procedure 20 times (including the random selection of 50 percent of the circuits), each time adding one year of data to the training sample and shifting the evaluation sample accordingly. We repeat the whole exercise 5 times, for a total of 100 training and evaluation samples.

TABLE 1. Details of the training and evaluation samples in the out-of-sample prediction exercise.

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.[3] 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

[3]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.

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. (2011) (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.[4] 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 ϕ, β 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).

  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

[4] 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.

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 ϕ, β 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.
- **Test-based forward model selection (TBFMS).** As an alternative forward model selection procedure, we also experiment with the test-based method proposed by Kozbur (2020). This class of algorithms selects covariates of progressively

larger-scale models and determines model size based on the outcome of statistical hypothesis tests. Following Kozbur (2020), we consider four versions of this method: (i) TBFMS-I, based on hypothesis tests for heteroskedastic disturbances; (ii) TBFMS-II, based on simplified hypothesis tests for heteroskedastic disturbances; (iii) TBFMS-III, based on fit-streamlined hypothesis tests for heteroskedastic disturbances; and (iv) TBFMS-IV, based on hypothesis tests for homoskedastic disturbances. The predictive density and point forecast of y_{T+1} implied by these models are constructed as in the post-lasso case.

References

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

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.

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.

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.

Kozbur, D. (2020): "Analysis of Testing-Based Forward Model Selection," *Econometrica*, 88, 2147–2173.

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.

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

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

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

Amazon and CEPR

Email address: dgiannon2@gmail.com

European Central Bank and ECARES

Email address: michele.lenza@ecb.europa.eu

Northwestern University, CEPR and NBER

Email address: g-primiceri@northwestern.edu