Macroeconomic forecast accuracy in a data-rich environment*

Maxime Leroux† Rachidi Kotchoni‡ Dalibor Stevanovic §

This version: February 2, 2018

Abstract

This paper proposes a new regularized data-rich model averaging forecasting technique. The performance of six classes of models at forecasting different types of economic series is evaluated in an extensive pseudo out-of-sample exercise. Our findings can be summarized in a few points: (i) Regularized Data-Rich Model Averaging techniques are hard to beat in general and are the best to forecast real variables. Simulations results show that this robust performance is attributable to the combination of sparsity/regularization with model averaging. (ii) The ARMA(1,1) model emerges as the best to forecast inflation growth, except during recessions. (iii) SP500 returns are predictable by data-rich models and model averaging techniques, especially during recessions. Also, factor models have significant predictive power for the signs of future returns. (iv) The forecast accuracy and the optimal structure of forecasting equations are quite unstable over time.

JEL Classification: C55, C32, E17

Keywords: Data-Rich Models, Factor Models, Forecasting, Model Averaging, Sparse Models, Regularization.

*We thank Mehmet Caner, Todd Clark, Marco Del Negro, John Galbraith, Eric Ghysels, Domenico Giannone, Serena Ng, Frank Schorfheide, participants of the Penn Big Data Conference and anonymous referees for valuable comments.

†Département des sciences économiques, Université du Québec à Montréal, 315, Ste-Catherine Est, Montréal, QC, H2X 3X2. Email: maxime.leroux246@gmail.com

‡Economix-CNRS, Université Paris Nanterre. Email: rachidi.kotchoni@u-paris10.fr

§Département des sciences économiques, Université du Québec à Montréal, 315, Ste-Catherine Est, Montréal, QC, H2X 3X2. Email: dstevanovic.econ@gmail.com. The author acknowledges financial support from the Fonds de recherche sur la société et la culture (Québec) and the Social Sciences and Humanities Research Council. Corresponding author.
Contents

1 Introduction ................................................. 1

2 Predictive Modeling ........................................ 4
   2.1 Standard Forecasting Models .......................... 4
   2.2 Data-Rich Models ....................................... 6
      2.2.1 Factor-Augmented Regressions ..................... 8
      2.2.2 Factor-Structure-Based Models ................... 11
      2.2.3 Data-Rich Model Averaging: the Complete Subset Regression (CSR) . . 13
      2.2.4 Regularized Data-Rich Model Averaging .......... 13
   2.3 Forecasts Combinations ............................... 14

3 Empirical Evaluation of the Forecasting Models ........ 15
   3.1 Data .................................................. 15
   3.2 Pseudo-Out-of-Sample Experiment Design ................ 16
   3.3 Variables of Interest .................................. 16
   3.4 Forecast Evaluation Metrics ........................... 17
      3.4.1 Point Forecast Evaluation ........................ 17
      3.4.2 Interval Forecast Evaluation ....................... 17
      3.4.3 Sign Forecast Evaluation ........................ 17
      3.4.4 Evaluation of Forecast Optimality ................ 18
   3.5 Model Confidence Set .................................. 18

4 Main Results ............................................... 19
   4.1 Industrial Production Growth .......................... 20
   4.2 Employment Growth .................................... 23
   4.3 CPI Inflation .......................................... 27
   4.4 Stock Market Index ..................................... 31

5 Stability of forecast accuracy ............................. 35
   5.1 Great Recession ....................................... 35
   5.2 Stability of Forecast Performance ...................... 38
   5.3 Stability of Forecast Relationships .................... 40

6 Simulation Evidence ....................................... 43

7 Conclusion ................................................. 45
1 Introduction

Many economic data sets have now reached tremendous sizes, both in terms of the number of variables and the number of observations. As all these series may not be relevant for a given forecasting exercise, one will have to preselect the most important candidate predictors according to economic theories, the relevant empirical literature and own heuristic arguments. In a Data-Rich environment, the econometrician is still left with a few hundreds of candidate predictors after this preselection process. Unfortunately, the performance of standard econometric models tends to deteriorate as the dimensionality of the data increases, which is the well-known curse of dimensionality. The new challenge is therefore to design computationally efficient methods capable of turning big datasets into concise information.1

When confronted with a large number of variables, econometricians often resort to sparse models, regularization or dense modeling. Sparse models involve a variable selection procedure that discards the least relevant predictors. In regularized (or penalized) models, a large number of variables are accommodated but a shrinkage technique is used to discipline the behavior of the parameters (e.g., LASSO, Ridge). LASSO type regularization leads to sparse models ex post as it constrains coefficients of least relevant variables to be null. In factor models, an example of dense modeling, the dynamics of a large number of variables is assumed to be governed by a small number of common components. All three approaches entail an implicit or explicit dimensionality reduction that is intended to control the overfitting risk and maximize the out-of-sample forecasting performance. In a recent study, Giannone et al. (2017) considered a Bayesian framework that balances the quest for sparsity with the desire to accommodate a large number of relevant predictors. They find that the posterior distribution of parameters is spreaded over all types of models rather than being concentrated on a single sparse model or a single dense model. This suggests that a well-designed model averaging technique can outperform any sparse model.

This paper proposes a new class of regularized data-rich model averaging techniques and contributes to the literature on predictive modeling of big data. Given the growing popularity of models that address big data issues, there is a need for an extensive study that compares their performance. This paper contributes to filling this gap by comparing the performance of six classes of models at forecasting the Industrial Production growth, the Employment growth, the Consumer Price Index acceleration (i.e., variations of inflation) and the SP500 returns.2 Only few studies have done such a comparison exercise. See Boivin and Ng (2005), Stock and

---

1Bayesian techniques developed in recent years to handle larger than usual VAR models can be viewed as an effort toward this objective. See Bambura et al. (2010), Koop (2013), Carriero et al. (2015) and Giannone et al. (2015), among others.

2These variables are selected for their popularity in the forecasting literature. Results for the Core CPI, interest rate and exchange rates variations are available in the supplementary material.

The first class of forecasting models considered consists of standard and univariate specifications, namely the Autoregressive Direct (ARD), the Autoregressive Iterative (ARI), the Autoregressive Moving Average ARMA(1,1) and the Autoregressive Distributed Lag (ADL) models. The second class of models consists of autoregressions that are augmented with exogenous factors: the Diffusion Indices (DI) of Stock and Watson (2002b), the Targeted DI of Bai and Ng (2008), the DI with dynamic factors of Forni et al. (2005) and the Three-pass Regression Filter (3PRF) of Kelly and Pruitt (2015). The third type of models assume that the factors are endogenous, meaning that the dynamics of the series being predicted obey the assumed factor structure. In the latter category, we have the Factor-Augmented VAR (FAVAR) of Boivin and Ng (2005), the Factor-Augmented VARMA (FAVARMA) of Dufour and Stevanovic (2013) and the Dynamic Factor Model (DFM) of Forni et al. (2005).

The fourth category consists of Data-Rich model averaging techniques known as Complete Subset Regressions (CSR) (see Elliott et al. (2013)). Our Regularized Data-Rich Model Averaging techniques are gathered in the fifth category. These are penalized versions of the CSR algorithm (CSR combined with preselection of variables or with Ridge regularization). Finally, the sixth category consists of methods that average all the available forecasts. Here we consider the naive average of all forecasts (AVRG), the median of all forecasts (MED), the trimmed average of all forecasts (T-AVRG) and the inversely proportional average of all forecasts (IP-AVRG). The latter forecasting method is considered in Stock and Watson (2004).

The data employed for this study are monthly macroeconomic series coming from McCracken and Ng (2015). The comparison of the models is based on their pseudo out-of-sample performance along five metrics: the Mean Square Prediction Error, the Mean Absolute Prediction Error, the ratio of correctly predicted signs, the coverage rate of an interval forecast and the p-value of a forecast optimality test à la Mincer-Zarnowitz. For each series, horizon and out-of-sample period, the hyperparameters of our models (number of lags, number of factors, etc.) are re-calibrated using the Bayesian Information Criterion (BIC). The variations of the optimal hyperparameters over time allows us to gage the stability of our forecast equations.

To the best of our knowledge, our paper is a rare attempt to put so many different models together and compare their predictive performance on several types of data in a pseudo out-of-sample forecasting experiment. Disentangling which type of models have significant forecasting power for real activity, prices and stock market is a valuable information for practitioners and policy makers. Another contribution of the current work is to provide a laboratory for

---

3CSR combined with LASSO penalty could have also been considered. However, the associated computational burden is prohibitive.

4For the sake of completeness, the simple random walk (RW) and the random walk with drift (RWD) are considered as well when relevant.
future development of forecasting models. The data used in this paper are publicly available and Matlab codes are available upon a simple request. The pseudo out-of-sample exercise generated a huge volume of empirical results. The presentation below will focus on highlights that convey the most important messages of the paper.

Irrespective of the forecast horizon and performance evaluation metrics, Regularized Data-Rich Model Averaging and Forecast Combinations techniques emerge as the best to forecast real variables. Factor Structure Based and Factor Augmented models are dominated in terms of Mean Square Prediction Error and Mean Absolute Prediction Error, but they are good benchmarks when the ratio of correctly predicted signs is considered. This is attributable to the fact that Data-Rich models involving factors are flexible enough to accommodate instability in the dynamics of the target, as suggested by Carrasco and Rossi (2016) and Pettenuzzo and Timmermann (2017). For the same reason, factor structure based and factor augmented models emerge among the best to forecast real variables during recessions.

The ARMA(1,1) emerges as an excellent parsimonious model to forecast the variations of inflation. This is in line with Stock and Watson (2007) and Faust and Wright (2013). One possible explanation for this good performance of the ARMA(1,1) is that inflation anticipations are well anchored so that its changes are exogenous with respect to the conditioning information set. Hence, Data-Rich models tend to be over-parameterized for this series and have poor predictive performance. Forecast combinations and Regularized Data-Rich Model Averaging compare favorably to the ARMA(1,1) at most horizons. During recessions, the ARMA(1,1) delivers its best performance three quarters ahead only, while model averaging and forecast combinations dominate at the other horizons.

In general, the best approaches to forecast the SP500 returns are Data-Rich Model Averaging and Forecast combinations. Factor Structure Based models have significant predictive power for the sign of the SP500 returns and even emerge as the best with respect to those metrics at long horizons. During recession, Data-Rich Model Averaging and Forecast combinations dominate at short horizon, while factor structure based models dominate at a longer horizon. The Random Walk (RW) model delivers the best coverage ratio for the SP500 returns three quarters ahead over the full out-of-sample period and six quarters ahead during recession periods. Abstracting from these exceptions, RW models are dominated with respect to all metrics and at all horizons. This suggests that stock returns are predictable to some extent.

Overall, our results show that sparsity and regularization can be smartly combined with model averaging to obtain a forecasting model that dominates state-of-the-art benchmarks. Our paper therefore provides a frequentist support for the conclusions found by Giannone et al. (2017) in their Bayesian framework. Another important finding is that the performance of models is not stable across the business cycle. More generally, we find overwhelming evidence of structural changes in all aspects of the forecasting equations.
Compared to previous literature, our results confirm that, when forecasting real activity series, substantial improvements over univariate time series models can be obtained by genius handling of large data sets. For instance, Stock and Watson (2006) have documented improvements up to 33% for industrial production, but our Regularized Data-Rich model averaging improves the MSPE between 29% and 48% with respect to the AR benchmark. Kim and Swanson (2014) found that the combination of factor modeling and shrinkage techniques works best in terms of MSPE while model averaging performs poorly. In contrast, our results suggest that data-rich model averaging combined with shrinkage techniques outperform in general other methods. However, mixing shrinkage and factor modeling à la Bai and Ng (2008) improves the forecasting performance during recessions. As for inflation growth, we show that simple ARMA(1,1) model is the best benchmark, while data-rich methods generally improves the forecasting precision during recessions. The presence of the MA component in inflation time series has been documented in few papers cited above, but the predictive performance of the ARMA model has not been shown in such a large horse race.

Finally, we compare the performance of the forecasting models based on data that are simulated using the data generating process (DGP) implied by a large-scale Dynamic Stochastic General Equilibrium (DSGE) model proposed by Ruge-Murcia and Onatski (2013). Given the high computational burden associated with this simulation exercise, we focus on two series (output growth and inflation growth) and three forecasting horizons (h = 1, 6 and 12). We find that our regularized data-rich model averaging techniques consistently achieve the best point and sign forecast performance when predicting output growth. Targeted CSR models are generally the best to predict inflation growth at short horizon while ARMA and ARI dominate at longer horizons. These results are in line with our empirical findings.

In the remainder of this papers we first present forecasting models in Section 2. Section 3 presents the design of the pseudo out-of-sample exercise. Section 4 reports the main empirical results. Section 5 analyzes the stability of the forecast accuracy. Section 6 presents simulation results and Section 7 concludes. Additional results are available in supplementary materials.

## 2 Predictive Modeling

This section presents the predictive models considered in this paper. Some of these predictive models are modifications of existing ones that are being tested for the first time.

### 2.1 Standard Forecasting Models

Let $Y_t$ denote a macroeconomic or financial time series of interest. If $\ln Y_t$ is a stationary process, we will consider forecasting its average over the period $[t+1, t+h]$ given by:
\[ y^{(h)}_{t+h} = \left( \frac{freq}{h} \right) \sum_{k=1}^{h} y_{t+k}, \]  

(1)

where \( y_t \equiv \ln Y_t \) and \( freq \) depends on the frequency of the data (e.g. 1200 if \( Y_t \) is monthly).

Most of the time, we are confronted with I(1) series in macroeconomics. For such series, our goal will be to forecast the average annualized growth rate over the period \([t+1, t+h] \), as in Stock and Watson (2002b) and McCracken and Ng (2015). We shall therefore define \( y^{(h)}_{t+h} \) as:

\[ y^{(h)}_{t+h} = \left( \frac{freq}{h} \right) \sum_{k=1}^{h} y_{t+k} = \left( \frac{freq}{h} \right) \ln \left( \frac{Y_{t+h}}{Y_t} \right), \]  

(2)

where \( y_t \equiv \ln Y_t - \ln Y_{t-1} \). In cases where \( \ln Y_t \) is better described by as an I(2) process, we define \( y^{(h)}_{t+h} \) as:

\[ y^{(h)}_{t+h} = \left( \frac{freq}{h} \right) \sum_{k=1}^{h} y_{t+k} = \left( \frac{freq}{h} \right) \left[ \ln \left( \frac{Y_{t+h}}{Y_{t+h-1}} \right) - \ln \left( \frac{Y_t}{Y_{t-1}} \right) \right], \]  

(3)

where \( y_t \equiv \ln Y_t - 2 \ln Y_{t-1} + \ln Y_{t-2} \).

Indeed, \( y^{(h)}_{t+h} \) is given by the same function of \( y_t \) everywhere while \( y_t \) is \( \ln Y_t \) in (1), the first difference of \( \ln Y_t \) in (2) and the second difference of \( \ln Y_t \) in (3). In the remainder of the section, we describe the standard forecasting models advocated in the paper.

**Autoregressive Direct (ARD)** Our first univariate model is the so-called *autoregressive direct* (ARD) model, which is specified as:

\[ y^{(h)}_{t+h} = \alpha^{(h)} + \sum_{l=1}^{L} \rho_l^{(h)} y_{t+l-1} + e_{t+h}, \quad t = 1, \ldots, T, \]  

(4)

where \( h \geq 1 \) and \( L \geq 1 \). A direct prediction of \( y^{(h)}_{T+h} \) is deduced from the model above as follows:

\[ \hat{y}^{(h)}_{T+h} = \hat{\alpha}^{(h)}(y_{T+h} - \hat{\alpha}^{(h)}) + \sum_{l=1}^{L} \hat{\rho}_l^{(h)} y_{T+1}, \]

where \( \hat{\alpha}^{(h)} \) and \( \hat{\rho}_l^{(h)} \) are OLS estimators of \( \alpha^{(h)} \) and \( \rho_l^{(h)} \). The optimal \( L \) will be selected using the Bayesian Information Criterion (BIC) for every out-of-sample (OOS) period. This makes the forecasting model more flexible by allowing the optimal \( L \) to vary over the OOS period.

**Autoregressive Iterative (ARI)** Our second univariate model is a standard AR(L) model specified as:

\[ y_{t+1} = \alpha + \sum_{l=1}^{L} \rho_l y_{t+1-l} + e_{t+1}, \quad t = 1, \ldots, T. \]  

(5)
where \( L \geq 1 \). This model is termed autoregressive iterative (ARI) because \( \hat{y}_{T+h|T}^{(h)} \) must be deduced from recursive calculations of \( \hat{y}_{T+1|T}, \hat{y}_{T+2|T}, \ldots, \hat{y}_{T+h|T} \).

We have:

\[
\hat{y}_{T+k|T} = \hat{\alpha} + \sum_{l=1}^{L} \hat{\rho} \hat{y}_{T+k-l|T}, \quad k = 1, \ldots, h,
\]

with the convention \( \hat{y}_{t|T} \equiv y_t \) for all \( t \leq T \) and:

\[
\hat{y}_{T+h|T}^{(h)} = (\text{freq}/h) \sum_{k=1}^{h} \hat{y}_{T+k|T}.
\] (6)

Equation (6) will remain the appropriate prediction formula for all iterative models as long as the definition of \( y_t \) is adapted to whether \( \ln Y_t \) is \( I(0), I(1) \) or \( I(2) \). The optimal lag \( L \) will be selected using the Bayesian Information Criterion (BIC) for every out-of-sample period.\(^5\)

**ARMA(1,1)** Dufour and Stevanovic (2013) showed that ARMA models arise naturally as the marginal univariate representation of observables when they jointly follow a dynamic factor model. This suggests that the ARMA(1,1) is a natural benchmark against which to evaluate the performance of Data-Rich models. The following representation is therefore considered and estimated by maximum likelihood:

\[
y_{t+1} = \alpha + \rho y_t + \theta e_t + e_{t+1}.
\] (7)

The prediction of \( y_{T+h} \) for any horizon \( h \) is computed using the formula (6) along with the output of the following recursion:

\[
\hat{y}_{T+k|T} = \hat{\alpha} + \hat{\rho} \hat{y}_{T+k-1|T} + \hat{\theta} \hat{e}_{T+k-1|T}, \quad k = 1, \ldots, h,
\]

where \( \hat{y}_{T|T} = y_T, \hat{e}_{T|T} = \hat{e}_T \) and \( \hat{e}_{T+k|T} = 0 \) for all \( k = 1, \ldots, h \).

**Autoregressive Distributed Lag (ADL)** A simple extension of the ARD model is obtained by adding exogenous predictors \( Z_t \) to its right-hand side. This leads to the so-called ADL model given by:

\[
y_{t+h}^{(h)} = \alpha^{(h)} + \sum_{l=1}^{L} \rho_l^{(h)} y_{t-l+1}^{(h)} + \sum_{k=1}^{K} Z_{t-k+1}^{(h)} \beta_k^{(h)} + e_{t+h},
\] (8)

where \( Z_t \) contains a small number of selected series. The precise content of \( Z_t \) is discussed in the empirical section.

\(^5\)The iterative approach is found to be better when a true AR(L) process prevails for \( y_t \) while the direct approach is more robust to misspecification, see Chevillon (2007). Marcellino et al. (2006) conclude that the direct approach provides slightly better results but does not dominate uniformly across time and series.
2.2 Data-Rich Models

There is a growing literature on how to deal with a large number of predictors when forecasting macroeconomic time series. The factor-based approaches started with the diffusion indices model of Stock and Watson (2002a) and Stock and Watson (2002b). Since then, several modifications and extensions of this model have been proposed.

Let $X_t$ be an $N$-dimensional stationary stochastic process. We consider a general DFM representation of $X_t$ that will serve as a basis for subsequent analyses. Following the notation of Dufour and Stevanovic (2013) and Stock and Watson (2005), we assume that:

$$X_t = \lambda(L) f_t + u_t,$$

$$u_t = \delta(L) u_{t-1} + \nu_t,$$

$$f_t = \gamma(L) f_{t-1} + \theta(L) \eta_t, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $f_t$ is a $q \times 1$ vector of latent common factors, $u_t$ is a $N \times 1$ vector of idiosyncratic components, $\nu_t$ is a $N \times 1$ vector of white noise that is uncorrelated with the $q \times 1$ vector of white noise $\eta_t$, $\lambda(L)$, $\delta(L)$, $\gamma(L)$ and $\theta(L)$ are matrices of lag polynomials. We have:

$$\lambda(L) = \sum_{k=0}^{p_\lambda-1} \lambda_k L^k,$$

$$\delta(L) = \sum_{k=0}^{p_\delta-1} \delta_k L^k,$$

$$\gamma(L) = \sum_{k=0}^{p_\gamma-1} \gamma_k L^k,$$

$$\theta(L) = I_q - \sum_{k=1}^{p_\theta} \theta_k L^k,$$

with $p_\lambda$, $p_\delta$, $p_\gamma$, $p_\theta \geq 1$ are the highest degrees of polynomials in each matrix. Indeed, the matrices of coefficients $\lambda_k$, $\delta_k$, $\gamma_k$ and $\theta_k$ are allowed to become sparse as $k$ increases to the maximum degrees so that the orders of the polynomials in a given matrix may vary.

For instance, the $i^{th}$ element of $X_t$ is represented as:

$$X_{it} = \sum_{k=0}^{p_\lambda-1} \lambda_{k,i} f_{t-k} + u_{i,t} \equiv \lambda^{(i)}(L) f_t + u_{it},$$

$$u_{it} = \sum_{k=0}^{p_\delta-1} \delta_{k,i} u_{i,t-1-k} + \nu_{it} \equiv \delta^{(i)}(L) u_{i,t-1} + \nu_{it},$$

where $\lambda_{k,i}$ is the $i^{th}$ row of $\lambda_k$, $\lambda^{(i)}(L) = \sum_{k=0}^{p_\lambda-1} \lambda_{k,i} L^k$, $\delta_{k,i}$ is the $i^{th}$ row of $\delta_k$ and $\delta^{(i)}(L) = \sum_{k=0}^{q_\delta-1} \delta_{k,i} L^k$.

We assume the approximate DFM by allowing for some limited cross-section correlations among the idiosyncratic components, see technical details in Stock and Watson (2005) and Bai 7.
and Ng (2008). The idiosyncratic errors \( \nu_{it} \) are assumed uncorrelated with the factors \( f_t \) at all leads and lags.

To obtain the static factor representation, we define \( F_t = [f'_t, f'_{t-1}, \ldots, f'_{t-p_{\lambda}+1}]' \), a vector of size \( K = qp_{\lambda} \) such that:

\[
X_t = \Lambda F_t + u_t, \quad (14)
\]
\[
u_t = \delta(L)u_{t-1} + \nu_t, \quad (15)
\]
\[
F_t = \Gamma F_{t-1} + \Theta(L)\eta_t, \quad (16)
\]

where

\[
\Lambda = \begin{bmatrix}
\lambda_0 & \lambda_1 & \ldots & \lambda_{p_{\lambda}-1}
\end{bmatrix}
\]
\[
\Gamma(L) = \begin{bmatrix}
\gamma_0 & \gamma_1 & \ldots & \gamma_{p_{\lambda}-1} \\
0 & I & 0 & \ldots \\
\vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 0 & I
\end{bmatrix} \quad \text{and} \quad \Theta(L) = \begin{bmatrix}
\theta(L) \\
0 \\
\vdots \\
0
\end{bmatrix}.
\]

Equations (14)-(16) define the FAVARMA model proposed in Dufour and Stevanovic (2013). A simplified version of this model where \( p_{\lambda} = 1 \) (so that \( K = q \) and \( \Theta(L) = \theta(L) \)) has been used in Bedock and Stevanovic (2016) to estimate the effects of credit shocks. A similar model with \( \theta(L) = I_q \) has been used to forecast time series in Boivin and Ng (2005) and to study the impact of monetary policy shocks in Bernanke et al. (2005).

In practice, \( q \) and \( p_{\lambda} \) cannot be separately identified due to the latent nature of \( f \). Therefore, we shall rewrite (16) in the static representation as a standard \( K \)-dimensional VARMA with no particular structure imposed on the matrices of coefficients. We have:

\[
F_t = \Phi(L)F_{t-1} + \Theta(L)\eta_t, \quad (17)
\]

where \( \Phi(L) = \sum_{k=0}^{p_{\phi}-1} \phi_k L^k \) and \( \Theta(L) \) is redefined as \( \Theta(L) = \sum_{k=0}^{p_{\theta}-1} \theta_k L^k \). The optimal values of \( p_{\phi} \) and \( p_{\theta} \) can be selected by BIC.

### 2.2.1 Factor-Augmented Regressions

The first category of forecasting models considered below are the factor-augmented regressions, where an autoregressive direct model is augmented with estimated static factors. In these models, there is no need to specify the dynamics of the factors as in (16) because static factors are extracted by principal component analysis. The second category of models are more directly related to the DFM model presented previously.
**Diffusion Indices (ARDI)** The first model is the (direct) autoregression augmented with diffusion indices from Stock and Watson (2002b):

\[
y_{t+h}^{(h)} = \alpha^{(h)} + \sum_{l=1}^{p^h_y} \rho^{(h)}_l y_{t-l+1} + \sum_{l=1}^{p^h_f} F_{t-l+1} \beta^{(h)}_l + e_{t+h}, \quad t = 1, \ldots, T
\]

\[
X_t = \Lambda F_t + u_t
\]

where \(F_t\) are \(K^{(h)}\) consecutive static factors and the superscript \(h\) stands for the value of \(K\) when forecasting \(h\) periods ahead. This means that the number of factors to be included in the predictive regression might differ across target variables and forecasting horizons. The optimal values of \(p^h_y\), \(p^h_f\) and \(K^{(h)}\) are simultaneously selected by BIC. The \(h\)-step ahead forecast is obtained as:

\[
y_{T+h|T}^{(h)} = \hat{\alpha}^{(h)} + \sum_{l=1}^{p^h_y} \hat{\rho}^{(h)}_l y_{T-l+1} + \sum_{l=1}^{p^h_f} F_{T-l+1} \hat{\beta}^{(h)}_l.
\]

The feasible ARDI model is obtained after estimating \(F_t\) as the first \(K^{(h)}\) principal components of \(X_t\). See Stock and Watson (2002a) for technical details on the estimation of \(F_t\) as well as their asymptotic properties. Below, we consider two variations of the ARDI model. In the first version, we select only a subset of \(K^{(h)}\) factors to be included in (18) while in the second the \(F_t\) are obtained as dynamic principal components.

**Variation I: ARDI-tstat** The importance of the factors as predictors of \(y_{t+h}^{(h)}\) may be independent of their importance as principal components. Indeed, the ordering of the factors in \(F_t\) is related to their capacity to explain the (co-)variations in \(X_t\). The selection of factors into the ARDI model automatically includes the first \(K^{(h)}\) principal components. A natural variation of this approach is to select only those that have significant coefficients in the regression (18). This leads to forecast \(y_{t+h}^{(h)}\) as:

\[
y_{T+h|T}^{(h)} = \hat{\alpha}^{(h)} + \sum_{l=1}^{p^h_y} \hat{\rho}^{(h)}_l y_{T-l+1} + \sum_{i \in K^*} \hat{F}_{i,T} \hat{\beta}^{(h)}_i
\]

\[
K^* = \{i \in 1, \ldots, K \mid t_i > t_c\}.
\]

where \(K^* \in K\) refers to elements of \(F_t\) corresponding to coefficients \(\beta^h_i\) having their t-stat larger (in absolute terms) than the critical value \(t_c\) (here we omit the superscript \(h\) for simplicity). Another difference with respect to the ARDI model is that the optimal number of factors changes over time.
Variation II: ARDI-DU  The second variation of the ARDI model is taken from Boivin and Ng (2005). The model is the same as the ARDI except that $F_t$ is estimated by one-sided generalized principal components as in Forni et al. (2005). Hence, the working hypothesis behind the dimensionality reduction is the DFM equation (9).

Targeted Diffusion Indices (ARDIT)  Another critique of the ARDI model is that not necessarily all series in $X_t$ are equally important to predict $y_{t+h}$. The ARDIT model of Bai and Ng (2008) takes this aspect into account. Instead of shrinking the factors space as in ARDI-tstat variation, the idea is first to pre-select a subset $X_t^*$ of the series in $X_t$ that are relevant for forecasting $y_{t+h}$ and next predict the factors using this subset. Bai and Ng (2008) propose two ways to construct the subset $X_t^*$:

- **Hard threshold (OLS):** ARDIT-hard
  
  $$
  y_{t+h} = \alpha^{(h)} + \sum_{j=0}^{3} \beta_j^{(h)} y_{t-j} + \beta_i^{(h)} X_{i,t} + \epsilon_t \tag{21}
  $$
  $$
  X_t^* = \{X_i \in X_t \mid t_{X_i} > t_c\} \tag{22}
  $$

- **Soft threshold (LASSO):** ARDIT-soft
  
  $$
  \hat{\beta}_\text{lasso} = \arg\min_\beta \left[ RSS + \lambda \sum_{i=1}^{N} |\beta_i| \right] \tag{23}
  $$
  $$
  X_t^* = \{X_i \in X_t \mid \beta_i^{\text{lasso}} \neq 0\} \tag{24}
  $$

In the hard threshold case, a univariate regression (21) is performed for each predictor $X_{it}$ at the time. The subset $X_t^*$ is then obtained by gathering those series whose coefficients $\beta_i^{(h)}$ have their t-stat larger than the critical value $t_c$. We follow Bai and Ng (2008) and consider 3 lags of $y_t$ in (21), and set $t_c$ to 1.28 and 1.65. The second approach uses the LASSO technique to select $X_t^*$ by regressing $y_{t+h}^{(h)}$ on all elements of $X_t$ and using LASSO penalty to discard uninformative predictors. As in Bai and Ng (2008) we target 30 series.

Three-Pass Regression Filter (3PRF)  Kelly and Pruitt (2015) propose another approach to construct predicting factors from a large data set. The factors approximation is in the spirit of the Fama-MacBeth two-step procedure:

1. Time series regression of $X_{it}$ on $Z_t$ for $i = 1, \ldots, N$

  $$
  X_{i,t} = \phi_{0,i} + Z_i^t \phi_i + \varepsilon_{i,t}
  $$
2. Cross-section regression of \( X_{it} \) on \( \hat{\phi}_i \) for \( t = 1, \ldots, T \)

\[
X_{i,t} = \varsigma_{0,t} + \hat{\phi}_i f_t + \epsilon_{i,t}
\]

3. Time series regression of \( y_{t+h}^{(h)} \) on \( \hat{f}_t \)

\[
y_{t+h}^{(h)} = \beta_0 + \beta \hat{f}_t + \eta_{t+h}
\]

4. Prediction

\[
\hat{y}_{T+h|T}^{(h)} = \beta_0 + \hat{\beta} \hat{f}_T
\]

We follow Kelly and Pruitt (2015) and use 4 lags of \( y_t \) as proxies for \( Z_t \). They also suggest an information criterion to optimally select the proxy variables.

2.2.2 Factor-Structure-Based Models

The second category of forecasting models relies directly on the factor structure when predicting the series of interest. The working hypothesis will be the DFM (9)-(11) or its static form (SFM) (14)- (16) with some variations. Another important difference is that the series of interest, \( y_t \), is now included in the informational set \( X_t \).

**Factor-Augmented VAR (FAVAR)** Suppose that \( X_t \) obeys the SFM representation (14)-(16) with \( \Theta(L) = \theta(L) = I \). We have:

\[
\begin{align*}
X_t &= \Lambda F_t + u_t \\
u_t &= \delta(L)u_{t-1} + v_t \\
F_t &= \Phi F_{t-1} + \eta_t.
\end{align*}
\]

This model implicitly assumes that \( p_\lambda = 1 \) so that \( K = q \) and \( F_t \) reduces to a first order VAR. The optimal order of the polynomial \( \delta(L) \) is selected with BIC while the optimal number of static factors is chosen by Bai and Ng (2002) \( IC_{q2} \) criterion. After estimation, one forecasts the factors using (27) upon assuming stationarity. The idiosyncratic component is predicted using (26) and then \( \hat{F}_t \) and \( \hat{u}_t \) are combined into (25) to obtain a prediction of \( X_t \). Boivin and Ng (2005) compare the direct and iterative approaches:

- Iterative

\[
\begin{align*}
\hat{F}_{T+h|T} &= \hat{\Phi} \hat{F}_{T+h-1|T} \\
\hat{u}_{T+h|T} &= \hat{\delta}(L) \hat{u}_{T+h-1|T} \\
\hat{X}_{T+h|T} &= \hat{\Lambda} \hat{F}_{T+h|T} + \hat{u}_{T+h|T}
\end{align*}
\]
The forecast of interest, \( \hat{y}_{T+h|T} \), is then extracted from \( \hat{X}_{T+h|T} \) or \( \hat{X}_{T+h|T}^{(h)} \). The accuracy of the predictions depends on the validity of the restrictions imposed by the factor model. As ARDI type models are simple predictive regressions, they are likely to be more robust to misspecification than the factor model.

**Factor-Augmented VARMA (FAVARMA)** Dufour and Stevanovic (2013) show that the dynamics of the factors should be modeled as a VARMA and suggest the class of Factor-Augmented VARMA models represented in (14)-(16). Since the VARMA representation is not identified in general, they suggest four identified forms of Equation (16): Final AR (FAR), Final MA (FMA), Diagonal AR (DAR) and Diagonal MA (DMA). Only the iterative version is considered:

\[
\begin{align*}
\hat{F}_{T+h|T} &= \hat{\Phi}\hat{F}_{T+h-1|T} + \sum_{k=1}^{p_\theta} \hat{\theta}_k \hat{\eta}_{T+h-k|T} \\
\hat{u}_{T+h|T} &= \hat{\delta}(L)\hat{u}_{T+h-1|T} \\
\hat{X}_{T+h|T} &= \hat{\Lambda}\hat{F}_{T+h|T} + \hat{u}_{T+h|T}
\end{align*}
\]

with \( \hat{\eta}_{T+h-k|T} = 0 \) if \( h - k > 0 \). The forecast \( \hat{y}_{T+h|T}^{(h)} \) is extracted from \( \hat{X}_{T+h|T} \).

**DFM** Contrary to the FAVAR(MA) approach, Forni et al. (2005) propose to use a nonparametric estimate of the common component to forecast the series of interest.\(^6\) The forecasting formula for the idiosyncratic component remains the same. The forecast of \( X_t \) is constructed as follows:

\[
\begin{align*}
\hat{u}_{T+h|T} &= \hat{\delta}(L)\hat{u}_{T+h-1|T} \\
\hat{X}_{T+h|T} &= \hat{\lambda}(L)\hat{f}_{T+h|T} + \hat{u}_{T+h|T}
\end{align*}
\]

and \( \hat{y}_{T+h|T}^{(h)} \) is extracted from \( \hat{X}_{T+h|T} \). The number of underlying dynamic factors \( f_t \) is selected by Hallin and Liska (2007)’s test. The advantage of the current approach over the FAVAR(MA) clearly lies in the nonparametric treatment of the common component, which might be more robust to misspecifications. However, the nonparametric method may struggle in finite samples.

\(^6\)See Boivin and Ng (2005) for discussion. It is the ‘DN’ specification in their paper.
2.2.3 Data-Rich Model Averaging: the Complete Subset Regression (CSR)

Unlike in the previous Data-Rich models, Elliott et al. (2013) do not assume a factor structure for the data. Instead, they propose to compute a large number of forecasts of $y_{T+h|T}$ using regression models that are based on different subsets of predictors in $X_t$. The final forecast is then obtained as the average of the individual forecasts:

$$y_{t+h,m}^{(h)} = c + \rho y_t + \beta X_{t,m} + \varepsilon_{t,m}$$  \hspace{1cm} (28)

$$y_{T+h|T}^{(h)} = \frac{\sum_{m=1}^{M} y_{T+h|T,m}^{(h)}}{M}$$  \hspace{1cm} (29)

where $X_{t,m}$ contains $L$ series for each model $m = 1, \ldots, M$. In Elliott et al. (2013) $L$ is set to 1, 10 or 20 and $M$ is the total number of models considered (up to 20,000 in specific cases). This method can be computationally demanding when the number of predictors in $X_t$ is large.

2.2.4 Regularized Data-Rich Model Averaging

In this section we consider the standard Lasso technique (which leads to sparse models) and two modifications of the Complete subset regression that build on the intuition of Giannone et al. (2017).

**Lasso** Here, the variable of interest can be predicted directly from the first step in Bai and Ng (2008) soft threshold targeted indices:

$$y_{t+h}^{(h)} = \alpha^{(h)} + \sum_{j=0}^{3} \beta_j^{(h)} y_{t-j} + \beta_i^{(h)} X_{i,t} + \epsilon_{t+h}$$

$$\hat{\beta}_{lasso} = \text{arg min}_{\beta} \left[ RSS + \lambda \sum_{i=1}^{N} |\beta_i| \right]$$

As suggested by Bai and Ng (2008), we tune the regularization parameter so as to select approximately 30 regressors. This is approximately the number of series that Giannone et al. (2017) found to be optimal.

**Targeted CSR** In the Targeted CSR, we preselect a subset of relevant predictors (first step) before applying the CSR algorithm (second step). This first step is meant to discipline the behavior of the CSR algorithm ex ante.

1. Soft or Hard Thresholding $\rightarrow X_t^* \in X_t$

2. CSR on $X_t^*$
Following Bai and Ng (2008), we reduce the set of predictors \( X \) into a subset \( X_t^* \) at the first step either by soft or hard thresholding. We consider four different specifications of Targeted CSR: soft thresholding with 10 and 20 regressors, and hard thresholding with 10 and 20 regressors.

**Ridge CSR** Alternatively, one may choose to use the entire set of predictors \( X \) but discipline the CSR algorithm ex post using a Ridge penalization. Each CSR predictive regression is estimated as follows

\[
\hat{y}_{t+h,m} = c + \rho y_t + \beta X_{t,m} \\
\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left[ RSS + \lambda \sum_{i=1}^{N} \beta_i^2 \right],
\]

and then the final forecast is constructed as usual:

\[
\hat{y}_{T+h|T} = \frac{\sum_{m=1}^{M} \hat{y}_{T+h|m}}{M}
\]

The intuition here is rather simple. As the CSR consists of combining a large number of forecasts obtained from randomly selected subsets of predictors, some subsets of predictors will likely be subject to multicolinearity problems. This is particularly an issue for macroeconomic application where many series are known to be highly correlated. A Ridge penalization permits to elude this problem and produces a well-behaved forecast from every subsample. We consider two specifications of Ridge CSR: one based on 10 regressors and another on 20 regressors.

### 2.3 Forecasts Combinations

Instead of looking at individual forecasts, one can also aggregate them into a single prediction.

**Equal-Weighted Forecast (AVRG)** The simplest, but often very robust, method is to set equal weights on each individual forecast, \( w_{it} = \frac{1}{M} \), i.e. take a simple average over all forecasts:

\[
y_{t+h|t}^{(h,ew)} = \frac{1}{M} \sum_{i=1}^{M} y_{t+h|i}
\]

**Trimmed Average (T-AVRG)** Another approach consists of removing the most extreme forecasts. First, order the \( M \) forecasts from the lowest to the highest value.
\((y_{t+h|t}^{(h,1)} \leq y_{t+h|t}^{(h,2)} \ldots \leq y_{t+h|t}^{(h,M)})\). Then trim a proportion \(\lambda\) of forecasts from both sides:

\[
y_{t+h|t}^{(h,\text{trim})} = \frac{1}{M(1-2\lambda)} \sum_{i=[\lambda M]}^{(1-\lambda)M} y_{t+h|t}^{(h,i)}
\]

where \(\lfloor \lambda M \rfloor\) is the integer immediately larger than \(\lambda M\) and \(\lfloor (1 - \lambda)M \rfloor\) is the integer immediately smaller than \((1 - \lambda)M\).

**Inversely Proportional Average (IP-AVRG)** A more flexible solution is to produce weights that depend inversely on the historical performance of individual forecasts as in Diebold and Pauly (1987). Here, we follow Stock and Watson (2004) and define the discounted weight on the \(i\)th forecast as follows

\[
w_{it} = \frac{m_{it}^{-1}}{\sum_{j=1}^{M} m_{jt}^{-1}},
\]

where \(m_{it}\) is the discounted MSPE for the forecast \(i\):

\[
m_{it} = \sum_{s=T_0}^{t-h} \rho^{t-h-s} (y_{s+h} - y_{s+h|s}^{(h,i)})^2,
\]

and \(\rho\) is a discount factor. In our applications, we consider \(\rho = 1\) and \(\rho = 0.95\).

**Median** Finally, instead of averaging forecasts one can use the median, another measure of central location, that is less subject to extreme values than the mean:

\[
y_{t+h|t}^{(h,\text{median})} = \text{median}(y_{t+h|t}^{(h,i)})_{i=1}^{M}
\]

The median further avoids the dilemma regarding which proportion of forecasts to trim.

### 3 Empirical Evaluation of the Forecasting Models

This section presents the data and the design of the pseudo-of-sample experiment.

#### 3.1 Data

We use historical data to evaluate and compare the performance of all the forecasting models described previously. The data employed consists of an updated version of Stock and Watson macroeconomic panel available at Federal Reserve of St-Louis’s web site (FRED). It contains
134 monthly macroeconomic and financial indicators observed from 1960M01 to 2014M12. Details on the construction of the series can be found in McCracken and Ng (2015).

The empirical exercise is easier when the data set is balanced. In practice, there is usually a trade-off between the relevance and the availability (and frequency) of a time series. Not all series are available from the starting date 1960M01 in the McCracken and Ng (2015) database, but this can be accommodated in the rolling window setup since a series that is not available at the starting date will eventually appear in the informational set as the window moves forward.

Our models all assume that the variables \( y_t \) and \( X_t \) are stationary. However, most macroeconomic and financial indicators must undergo some transformation in order to achieve stationarity. This suggests that unit root tests must be performed before knowing the exact transformation to use for a particular series. The unit root literature provides much evidence on the lack of power of unit root test procedures in finite samples, especially with highly persistent series. Therefore, we simply follow McCracken and Ng (2015) and Stock and Watson (2002b) and assume that price indexes are all \( I(2) \) while interest and unemployment rates are \( I(1) \).  

3.2 Pseudo-Out-of-Sample Experiment Design

The pseudo-out-of-sample period is 1970M01 - 2014M12. The forecasting horizons considered are 1 to 12 months. There are 540 evaluation periods for each horizon. All models are estimated on rolling windows. We have compared the forecast accuracy of rolling versus expanding (or recursive) windows, and the results are similar. For each model, the optimal hyperparameters (number of factors, number of lags, etc.) are selected specifically for each evaluation period and forecasting horizon. The size of the rolling window is \( 120 - h \) months.

3.3 Variables of Interest

We focus on four variables in the subsequent presentation: Industrial Production (INDPRO), Employment (EMP), Consumer Price Index (CPI) and SP500 index. INDPRO and EMP are real activity variables, CPI is a nominal variable while the SP500 represents the stock market. Additional results are available in the supplementary material for the Core Consumer Price Index (Core CPI), the 10-year treasury constant maturity rate (GS10) and the US-UK and US-Canada bilateral exchange rates. The logarithm of the real series (INDPRO and EMP) and

---

7Bernanke et al. (2005) keep inflation, interest and unemployment rates in levels. Choosing (SW) or (BBE) transformations has effects on correlation patterns in \( X_t \). Under (BBE), the group of interest rates is highly correlated as well as the inflation rates. As pointed out by Boivin and Ng (2006), the presence of these clusters may alter the estimation of common factors. Under (SW), these clusters are less important. Recently, Banerjee et al. (2014) and Barigozzi et al. (2016) proposed to deal with the unit root instead of differentiating the data.
the SP500 are treated as I(1) while the logarithm of the CPI is assumed to be I(2), as in Stock and Watson (2002b) and McCracken and Ng (2015).

3.4 Forecast Evaluation Metrics

The forecasting models will be compared using five metrics. Two of those evaluate the quality of point forecasts, one metric evaluates the quality of interval forecasts, one evaluates the quality of sign predictions and the last one assesses the forecast optimality à la Mincer-Zarnowitz.

3.4.1 Point Forecast Evaluation

Following a standard practice in the forecasting literature, we evaluate the quality of our point forecasts using the root Mean Square Prediction Error (MSPE) and the Mean Absolute Prediction Error (MAPE). We advocate these metrics ex post as ad hoc performance evaluation tools and do not attempt to relate them ex ante to a cost function at the model estimation stage. Indeed, the forecasting models are estimated using different algorithms, some of which are not directly related to the root MSPE or MAPE.

Although these metrics are interesting, they miss important aspects of the distribution of the forecasts. The next performance criterion addresses this drawback.

3.4.2 Interval Forecast Evaluation

Let \( \hat{y}_{t+h|t} \) be the point forecast and \( \sigma_h^2 \) the associated variance. Assuming normality for the forecasting errors leads to:

\[
y_{t+h} \sim N(\hat{y}_{t+h|t}, \sigma_h^2).
\]  
(32)

Hence, an interval forecast can be deduced as \( \hat{y}_{t+h|t} \pm c \times \sigma_h \) where \( c \) is selected for a given nominal coverage rate of the \((1 - \alpha)\%\).

In our empirical experiments, the metrics of interest is the actual coverage ratio of an out-of-sample interval forecast that has a nominal coverage rate of 70\%. Hence, \( \hat{y}_{t+h|t} \) and \( \sigma_h^2 \) are both estimated out-of-sample. We use the first 50 observations of the out-of-sample period to calculate the initial estimate of the error variance \( \sigma_h^2 \). This estimate is then updated recursively.

3.4.3 Sign Forecast Evaluation

Here, we compare the forecasting methods in terms of their ability to correctly predict the signs of the target series. Indeed, a forecasting model that is outperformed by the RW according to the MSPE or MAPE can still have significant predictive power for the sign of the target variable, see Satchell and Timmermann (1995). This possibility can be assessed by means of
the Pesaran and Timmermann (1992) sign forecast test. The test statistic is given by:

\[ S_n = \frac{\hat{p} - \hat{p}^*}{\sqrt{\text{Var}(\hat{p}) - \text{Var}(\hat{p}^*)}}, \]

where \( \hat{p} \) is the sample proportion of correctly signed forecasts (or the success ratio) and \( \hat{p}^* \) is the estimate of its expectation. This test statistic is not influenced by the distance between the realization and the forecast as is the case for MSPE or MAPE. Under the null hypothesis that the signs of the forecasts are independent of the signs of the target, we have \( S_n \rightarrow N(0, 1). \)

The ratio of correctly predicted signs are presented in the main text. The results of the formal significance tests for the predictive power of the forecasting models for the signs of the target variable are presented in supplementary materials.

### 3.4.4 Evaluation of Forecast Optimality

Finally, we compare the forecasting methods with respect to their optimality by means of Mincer-Zarnowitz regressions. For each forecasting model and horizon, the following regression is estimated:

\[ y_{t+h} = \beta_0 + \beta_1 \hat{y}_{t+h|t} + u_{t+h}, \tag{33} \]

where \( \hat{y}_{t+h|t} \) is an out-of-sample forecast of \( y_{t+h} \). If \( \hat{y}_{t+h|t} \) is optimal with respect to the information set on which the forecasting exercise is based, we should have \( (\beta_0, \beta_1) = (0, 1) \). Hence, a measure of forecast optimality can be obtained as the p-value of the test for the null hypothesis that \( (\beta_0, \beta_1) = (0, 1) \). This test must be conducted separately for every \( h \) or jointly for all horizons. See Patton and Timmermann (2012) who study this type of optimality testing. Below, the results of the optimality tests are presented separately for each horizon while the joint tests are deferred to the supplementary material.

### 3.5 Model Confidence Set

In addition to the performance metrics presented above, we also consider the Model Confidence Set (MCS) approach advocated in Hansen et al. (2011). This approach allows us to select a subset of best models, with a given level of confidence. This concept is of particular interest in our paper where we have 31 forecasting models in total. The MCS approach does not assume that a particular model is the true one. For a given performance metrics, the MCS is obtained by first finding the best forecasting model, and then selecting the models are not significantly different from the best model with a given confidence level.

---

8 Let \( q \) denote the proportion of positive realizations in the actual data and \( \hat{q} \) the proportion of positive forecasts. Under \( H_0 \), the estimated theoretical number of correctly signed forecast is \( \hat{p}^* = q\hat{q} + (1 - q)(1 - \hat{q}) \).
We construct the subsets of best models based on the quadratic loss function, for each target variable and by horizon. The results vary considerably with the level of confidence. As expected, the \((1 - \alpha)\) confidence intervals contain less models when \(\alpha\) is larger. Following Hansen et al. (2011), we present the empirical results for 75% confidence interval. Supplementary material contains results for \(\alpha = 10\%, \ 50\%\). The number of bootstrap replications is set to 4000.

4 Main Results

This section presents our main empirical results for the industrial production and employment growth rates, the variations of inflation and the returns on the SP500 index. In total, we have 31 forecasting models, of which 26 are individual forecasts and 5 are forecast combinations. The results are summarized in spider charts where each dimension represents a metric. This type of graphical representation is convenient as it allows us to present a large amount of results under space constraints. The first and second dimensions of the spider chart represent the Root MSPE (RMSPE) and Root MAPE (RMAPE) calculated as \([\max (C) - C_j]/(\max (C) - \min (C))\], where \(C_j\) is either the RMSPE or RMAPE of model \(j\) and \(\min (C)\) and \(\max (C)\) are taken over \(j\). The third and fourth dimensions are the percentage of correctly predicted signs (SR) and the empirical coverage ratio of interval forecasts with 70% nominal coverage (CR). The fifth dimension is the p-value of the Mincer Zarnowitz optimality test (MZ). The ideal value for these metrics is unity but the MZ metric is truncated to 0.15 for the sake of legibility. The farther a model lies from the origin of the chart along a given dimension, the better this model is for the corresponding metrics.

The figures are not legible when each model is identified by a different marker. Therefore, we have chosen to identify each family of model by a different marker. As a reminder, note that our models are partitioned into six groups: standard time series models, factor-augmented models, factor structure-based models, Data-Rich Model Averaging, Regularized Data-Rich Model Averaging and Forecast Combinations (see the legend of the spider charts). The best model along each dimension is indicated at the corresponding edge of the spider chart. The supplementary material contains tables and figures that present the empirical results in more details.\(^9\) The MCS results are summarized in few figures. They list all models and report which have been selected for a given horizon.

The analysis is done for the full out-of-sample period as well as for NBER recessions only. We use historical data and assume ex post that a recession has occurred at the periods identified by the NBER. Indeed, the knowledge of which models have performed best in the past during recessions is of interest for policy makers and practitioners, even for real-time decision making.

\(^9\)The tables with MSPE relative to the ARD (autoregressive direct) model used as a benchmark are available in the supplementary material, along with the Diebold and Mariano (1995) (DM) test.
If the probability of a recession is high enough at the current period, our result can be used to identify a subset of forecasting models that have been historically more accurate than the others during recessions.

4.1 Industrial Production Growth

In this section, we examine the performance of the various forecasting models for industrial production growth. Figure 1 present the results for the Full Out-of-sample period (1970-2014) while Figure 2 is restricted to NBER recessions periods (i.e., target observation belongs to a recession episode).

We note that Forecast Combinations and Regularized Data-Rich Model Averaging approaches dominate the others over the Full Out-of-sample period. This is true irrespective of the forecast horizon and the performance evaluation metrics. Some standard Data-Rich models averaging techniques are close to the envelope of the spider charts as well. More often than not, Factor Augmented and Factor Structure-based models dominate standard univariate models. However, the optimality of the forecasts of factor augmented models deteriorates as the horizon increases. Some factor structure based models are more resilient in that respect, especially at the horizon longer than $h = 1$. At the horizon $h = 3$, three different versions of the CSR dominate the other models in terms of RMSPE, RMAPE and SR while LASSO dominates in terms of CR.

During recessions, economic variables tend to change at a faster pace and uncertainty is higher than usual. As a result, models that are quite flexible perform well during these periods while standard time series models are largely dominated. Indeed, we find that some factor structure based and factor augmented models now emerge among the best. Forecast combinations and Regularized Data-Rich models averaging techniques still perform very well relatively to the best alternative benchmark along each dimension. The performance of most forecasting models worsens during recessions, in particular in terms of the SR, CR and MZ metrics. The optimality metrics (MZ) is the one that suffers the most during recessions. At the horizon $h = 3$, FAVARMA models dominate the other approaches in terms of RMSPE and CR during recessions while FAVARI model dominates in terms of RMAPE and SR.
Figure 1: Forecasting Industrial Production: Full OOS

Note: Each dimension in this spider chart represents an evaluation metrics. RMAPE and RMSPE stand for root mean absolute and squared predictive errors respectively. SR is the success ratio in sign prediction, CR is the coverage rate for interval forecasts. MZ represent the p-value of the forecast optimality test performed in Mincer-Zarnowitz regressions.
Figure 2: Forecasting Industrial Production: NBER Recession

- Standard Time Series
- Factor–augmented
- Factor–structure
- Data–Rich Model Avrg
- Regularized Data–Rich Model Avrg
- Fcsts Combinations
4.2 Employment Growth

We now examine the empirical results for Employment Growth. Figure 3 shows the results for the Full Out-of-sample period while Figure 4 focuses on data points that belong to NBER recessions. The results are quite similar to what is obtained for industrial production growth.

On the Full out-of-sample period, Regularized Data-Rich Models Averaging, and in particular the CSR with Ridge regressions of 20 predictors (CSR-R,20), emerge as the best techniques. Forecasts combinations are the best at very short horizon. Factor augmented and factor structure based models lag slightly behind in terms of RMSPE and RMAPE but they compare favorably to the best alternatives in terms of the other criteria. Data-Rich model averaging techniques are often dominated by their Regularized counterparts, especially in terms of RM-SPE and RMAPE. As previously, some Factor Structure based models perform very well in term of MZ. Standard time series models are less dominated than previously, especially at the horizon $h = 1$ and for the SR, CR and MZ metrics. At the horizon $h = 3$, versions of the CSR dominate in terms of RMSPE, RMAPE and SR while LASSO dominates in terms of CR.

Although quite resilient, forecast combinations are no longer the best techniques during recession episodes. They are dominated by factor augmented models, factor structure-based models and Regularized Data-Rich model averaging. Factor structure-based models perform better than in the previous section relatively to the other models at short horizon. Univariate time series models are not to be recommended during recessions. As argued in the previous section, univariate models are not flexible enough to capture the rapid changes in the dynamics of economic variables and in the structure of their mutual correlation that occur during recessions.

Figure 5 reports the MCS for industrial production and employment. We note that no univariate model is selected in the MCS while almost all Regularized Data-Rich and forecasts combinations are selected for the full out-of-sample. When the forecasting periods are restricted to NBER recessions, Factor-augmented regressions and FAVAR(MA)s are more frequently included, at the expense of forecasts combinations. Interestingly, Lasso is almost always in MCS during recessions while it is not selected at all when the full OOS period is considered.

In summary, our Regularized Data-Rich model averaging techniques and forecast combinations are the best techniques to predict real activity variables in general. During recessions, the Regularized Data-Rich model averaging, factor augmented and factor structure-based models dominate. This reflects the fact that the efficiency of the latter three approaches relatively to the other methods increases during recession periods. The LASSO emerges as a robust interval forecast technique of real activity variables during the full out-of-sample period.
Figure 3: Forecasting Employment: Full OOS

- Standard Time Series
- Factor-augmented
- Factor-structure
- Data-Rich Model Avrg
- Regularized Data-Rich Model Avrg
- Fcsts Combinations
Figure 4: Forecasting Employment: NBER Recession

- **h=1**
- **h=3**
- **h=6**
- **h=9**
- **h=12**

**RMSPE**

- **FAVARD**
- **RMAPE**
- **ARDI**
- **SR**
- **ADL**
- **CR**
- **ARMA(1,1)**
- **MZ**

**Standards**

- **Standard Time Series**
- **Factor–augmented**
- **Factor–structure**
- **Data–Rich Model Avrg**
- **Regularized Data–Rich Model Avrg**
- **Fcasts Combinations**
The figure shows 75% model confidence sets of best models to forecast the two real activity series, in terms of RMSPE. Circles represent industrial production, and Xs the employment results.
4.3 CPI Inflation

We now examine the performance of the various models at forecasting the variations of inflation deduced from the consumer price index (CPI). The target series of interest here is therefore the second difference of the logarithm of the CPI (i.e., CPI acceleration). Figure 6 shows the results for the entire out-of-sample period while Figure 7 is restricted to recession periods.

Over the whole out-of-sample period, the ARMA(1,1) surprisingly dominates all individual Data-Rich models at most forecasting horizons. For instance, the ARMA(1,1) dominates the other models in terms of RMSPE and RMAPE at the horizon $h = 3$.\textsuperscript{10} Forecast combinations are the second-best performing approaches at most horizon, followed by Regularized Data-Rich model averaging. At the horizon $h = 3$, factor augmented and factor structure-based models dominate in terms of the SR and CR metrics. At long forecast horizon ($h = 12$), factor augmented models and Regularized Data-Rich model averaging are more resilient than forecast combinations in terms of the MZ criterion.

During NBER recessions, the ARMA(1,1) model is dominated, except for one-quarter horizon in terms of RMSPE and RMAPE. Other approaches such as Forecast Combinations, ADL, Targeted Diffusion Indices or Regularized Data-Rich model averaging now dominate for all other horizons and metrics. One plausible explanation for the good performance of the ARMA(1,1) on the full out-of-sample period is that inflation is generally well anticipated so that its variations behave like an exogenous noise. Consequently, Data-Rich models tend to be over-parameterized and have poor predictive performance for this series. During recessions specifically, economic variables are subject to unusually large shocks while agents anticipations change rapidly over time and are quite noisy. As a result, the ARMA(1,1) model loses its predictive power and data-rich models become favored. An important lesson learned from these results is that a model that outperforms the others in terms of RMSPE can be dominated by another model at predicting the sign of the target variable. In case of inflation change, some Data-Rich models should be preferred to the ARMA(1,1) model.

Figure 8 reports the model confidence sets for CPI and Core CPI inflation growths. Compared to the MCS for real activity variables, fewer models are selected here. Over the full out-of-sample period, the ARMA(1,1) is always included, as well as CSR-R,\textsuperscript{10} and inversely proportional average. A few factor-augmented regressions are selected while almost no factor structure-based or CRS models are included in the MCS. The results are different during recessions where many factor-augmented regressions are selected in the MCS. The results are different during recessions where many factor-augmented regressions are selected in the MCS. Almost all forecasting methods are selected at short horizons (i.e., 1 to 3 months ahead).

\textsuperscript{10}Stock and Watson (2007) suggest that the MA part of the inflation stochastic process has increased from 1984. Ng and Perron (1996) and Ng and Perron (2001) also document similar evidence. Foroni et al. (2017) found that the MA part improves the forecasting power of mixed-frequency models when predicting the U.S. inflation.
Figure 6: Forecasting CPI Inflation: Full OOS

- Standard Time Series
- Factor–augmented
- Factor–structure
- Data–Rich Model Avrg
- Regularized Data–Rich Model Avrg
- Fcsts Combinations
Figure 7: Forecasting CPI Inflation: NBER Recession

For each forecast horizon $h$, the figure depicts the performance of different models, including:

- Standard Time Series
- Factor-augmented
- Factor-structure
- Data-Rich Model Avrg
- Regularized Data-Rich Model Avrg

The diagrams show the root mean squared error (RMSPE) and relative mean absolute percentage error (RMAPE) for various models. The following models are shown:

- ARMA(1,1)
- ADL
- ARDI-hard,1.65
- ARDI-hard,1.28

The models are evaluated at forecast horizons $h = 1, 3, 6, 9, 12$. The diagrams illustrate the relative performance of each model across different forecast horizons.
The figure shows 75% model confidence sets of best models to forecast the two inflation series, in terms of RMSPE. Circles represent CPI, and Xs the Core CPI results.
4.4 Stock Market Index

We now examine the empirical results for the SP500 returns. Ideally, the forecasting of stock market returns should be done using the real-time vintages of the predictors. Unfortunately, these vintages are not always available for a large number of series. Our results are therefore based on the latest available information on all variables. Figure 9 shows the results for the entire out-of-sample period while Figure 10 is restricted to recession periods.

Given the perpetual debate on the efficiency of stock markets and the predictability of stock returns, we have highlighted the performance of the random walk models (RW and RWD). These two models are dominated by several methods, during the full out-of-sample period. This clearly supports that stock returns are predictable to some extent. The RW model emerges as a good benchmark in terms of the CR metrics at the horizon $h = 3$. It is the best model in terms of the RMSPE and RMAPE during recessions at long horizon ($h=12$).

Over the full out-of-sample period, forecast combinations and Data-Rich model averaging (Regularized and dense) are the best performing forecasting techniques for the SP500 index. Factor structure-based and factor augmented models are in general dominated but they often emerge as good benchmarks when the SR metric is considered. Hence, an investor who is rather interested in predicting the direction of change of the stock market index would rather favor a factor models, especially at horizons $h = 9$ and beyond. Interestingly, our CSR-R model is the only model for which the forecast optimality is not rejected for longer horizons. The coverage rates are close to the nominal level for most of the models and horizons.

During recessions, forecast combination, Regularized Data-Rich model averaging and factor augmented models perform well at short horizon but their performance deteriorate in long run. Factor structure based models are slightly dominated at short horizons and more resilient at longer horizons ($h = 9$ and beyond). However, forecast optimality is largely rejected for all models for $h > 4$. The coverage rate heavily shrinks during recession due to volatility spikes on stock markets, but the sign prediction success rates are more resilient. Overall, factor structure based models deliver the most robust forecasts for the SP500 returns, by achieving a good balance between forecast accuracy during expansions versus during recessions, and also between the precision of the point forecast and that of the sign forecast.11

Figure 11 reports the MCS for the prediction of the SP500 returns. Many models are selected over the full out-of-sample. In particular, we note the presence of almost all univariate models as well as CSR, CSR-Ridge and forecasts combinations in the MCS. During NBER recessions, the selection is much sharper. The random walk is the sole model that is present in the MCS at all horizons, except $h = 1$. FAVAR(MA)s and some CSR-Ridge and forecasts combinations are also resilient, while the Lasso is included for horizons longer than 6 months.

11Rapach and Zhou (2010) also found that stock market returns can be predictable during recessions.
Figure 9: Forecasting SP500: Full OOS

- Standard Time Series
- Factor-augmented
- Factor-structure
- Data-Rich Model Avrg
- Regularized Data-Rich Model Avrg
- Fcst Combinations
- ** Random walks
Figure 10: Forecasting SP500: NBER Recession

- **h=1**
  - RMSPE
  - ARDI
  - RMAPE
  - ARDI
  - SR
  - ADL
  - CR
  - T−CSR−h,20
  - MZ

- **h=3**
  - RMSPE
  - ARDI
  - RMAPE
  - ARDI
  - SR
  - ADL
  - CR
  - T−CSR−h,20
  - MZ

- **h=6**
  - RMSPE
  - ARDI
  - RMAPE
  - ARDI
  - SR
  - ADL
  - CR
  - T−CSR−h,20
  - MZ

- **h=9**
  - RMSPE
  - ARDI
  - RMAPE
  - ARDI
  - SR
  - ADL
  - CR
  - T−CSR−h,20
  - MZ

- **h=12**
  - RMSPE
  - ARDI
  - RMAPE
  - ARDI
  - SR
  - ADL
  - CR
  - T−CSR−h,20
  - MZ

Legend:
- **Standard Time Series**
- **Factor-augmented**
- **Factor-structure**
- **Data-Rich Model Avrg**
- **Regularized Data-Rich Model Avrg**
- **Fcsts Combinations**
- **Random walks**
Figure 11: Forecasting SP500: MCS

The figure shows 75% model confidence sets of best models to forecast the stock prices, in terms of RMSPE.
5 Stability of forecast accuracy

In this section, we first closely look at the forecasting exercise during the Great Recession. Secondly, we study the stability of forecast performance over the out-of-sample period. Finally, we examine the stability of the factor-based forecasting equations over time.

5.1 Great Recession

Figure 12 plots the 3-month ahead out-of-sample forecasts of industrial production, employment, CPI and SP500 during the Great Recession. The most pessimistic forecasts (lowest percentiles of the cross-sectional distribution of forecasts) appear to be the best predictor of Industrial Production growth and Employment growth during most recessions. This suggests that in a real-life application, e.g. stress-testing for financial institutions, the pessimistic forecasts may be used as worse case scenarios that become more realistic on the eve of an economic crisis. Another interesting observation is that during recovery, the optimistic forecasts (high quantiles) closely follow industrial production but are too optimistic for the employment.

In the case of CPI inflation, ARMA forecasts track the realized values well most of the time while the low percentiles are too pessimistic from 2009 onward. The pessimistic scenarios predict downturns in stock markets, but the fast recovery at the end of the recession is not well predicted even by the highest percentile of the forecasts distribution.

Figure 13 plots the 3-month ahead out-of-sample 90% interval predictions of the best MSPE models as well as the distribution of all forecasts.\textsuperscript{12} In case of real activity series and the SP500 the interval forecast of a single model is usually much wider than the distribution of all point forecasts, except for CPI inflation. However, the dispersion in actual forecasts explodes during the Great Recession such that it becomes even larger than the 90% density forecast. This suggests that the empirical distribution of forecasts contains relevant information beyond the density prediction of the best MSPE model. It is particularly important around business cycle turning points where the lower percentiles and the dispersion of forecasts are quite informative.

\textsuperscript{12}Results are similar for 70% interval forecasts.
The figure shows the 3-month ahead pseudo-out-of-sample forecasts during the Great Recession. The blue line presents the historical data and the black line the forecast of the best MSPE model. The gray area around these lines presents the forecasts of all models. Other lines present the quantiles of the distribution of all forecasts.
The figure shows the 3-month ahead pseudo-out-of-sample 90% interval forecasts of the best MSPE models during the Great Recession. The dark grey area represents the distribution of all 31 forecasts.
5.2 Stability of Forecast Performance

Here we verify the stability of the forecast accuracy in our pseudo-out-of-sample exercise.\textsuperscript{13} Figures 14 plots the 3-year moving average of the root MSPE of selected models for 3-month predictions. There is a huge downturn in the level of MSPE for real activity series from middle '80s, except for the Great Recession period, which coincides with the Great Moderation period. The situation with CPI is different. The forecasting errors rise since 2000 and fly to historical peaks during the Great Recession. However, it dropped back to the usual level since then. The forecast errors of SP500 returns are closely related to NBER recession cycles. The comparison of forecast accuracy over the Great Moderation, through spider charts and MCS, is reported in supplementary material.

Figure 14: Root MSPE over time

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure14.png}
\caption{Root MSPE over time}
\end{figure}

The figure shows the 3-year moving average of the root MSPE of selected models for 3-month ahead horizon.

\textsuperscript{13}See Giacomini and Rossi (2009), Rossi and Sekhposyan (2010) and Rossi and Sekhposyan (2011), among others, for recent examples of the time-varying forecast performance.
Giacomini and Rossi (2010) propose a test to compare the out-of-sample forecasting performance of two competing models in the presence of instabilities. The idea is to test whether the forecasting errors are different *during* the out-of-sample period instead of looking only at the global performance as is usually done with the Diebold-Mariano test. Figure 15 shows the results of the Giacomini-Rossi fluctuation test for several horizons and two critical values. We report the comparison between the overall best MSPE model for each series and the ARD alternative. The moving average of the standardized difference of MSPEs is produced with 54-month window, which corresponds to 10% of the out-of-sample period. The results point to considerable instability in the forecast accuracy across horizons and over time.

Figure 15: Giacomini-Rossi fluctuation test

The figure shows the Giacomini-Rossi fluctuation test for best RMSPE models against the ARD benchmark. CV, 0.05 and CV, 0.10 correspond to 5% and 10% critical values respectively.
5.3 Stability of Forecast Relationships

Several recent studies have suggested that factor loadings and the number of factors are likely to change over time. The results from our exercise point in the same direction. The number of principal components retained in factor-augmented models vary considerably across the out-of-sample period as well as for different forecasting horizons and across the series of interest.

In general, forecasting real activity measures require more factors (and their lags) than when predicting inflation and stock returns. Figure 16 plots the number of series selected by soft (Lasso) and hard thresholds for all series at 3-month horizon. Recall that this is the first step in ARDIT models as well as in our targeted CSR model. The patterns of the two real activity series are quite similar. The number of candidate predictors is generally lower when predicting CPI inflation growth. In the case of stock returns the number of selected series is declining until the Great Recession.

Figure 17 shows the type of series selected by hard thresholding with $t_c = 1.65$ for 3-month ahead predicting. We group the data as in McCracken and Ng (2015) and show whether a series has been selected or not over the whole out-of-sample period. The picture shows that there is a lot of instability in the selection of variables. The probability that a particular predictor will be consistently selected is higher for some groups and depends on the series being predicted. For instance, several indicators in Employment & Hours, Consumption and Money & Credit groups are often present when predicting industrial production and employment. There is a lot of instability in predictor selection for CPI where only a small number of candidates are systematically present. Similar pattern is observed in case of SP500. Overall, our very long out-of-sample period and the variety of forecasting models may serve as a good laboratory to study the stability of factor structures and the forecasting relationships. The results presented in this section document the prevalence of structural changes in all dimensions. However, the occurrence of these changes are not evenly distributed across the forecasted series and forecasting horizons.

---


15 For the sake of space the figures are presented in the supplementary material.
Figure 16: Number of series pre-selected by hard and soft thresholding

The figure shows the number of series selected by the hard and soft thresholding when predicting at 3-month horizon.
The figure shows the series pre-selected by the hard thresholding with $t_c = 1.65$ when predicting at 3-month horizon. The content of each group is described in McCracken and Ng (2015).
6 Simulation Evidence

In order to verify the robustness of our empirical results, we have simulated artificial data from a model that is not loaded in favor of a particular forecasting technology. We use a DGP that is deduced from a large multi-sector Dynamic Stochastic General Equilibrium (DSGE) model proposed by Ruge-Murcia and Onatski (2013). This DSGE model is capable of generating 6 aggregate series and 150 disaggregate series. The calibration of the model is done using US data. Data are simulated from a linear state-space representation with three pervasive common dynamic shocks and 30 sectoral productivity shocks. Ruge-Murcia and Onatski (2013) showed that principal components can hardly replicate the common factor space, but diffusion indices do improve forecasts of aggregate output growth and inflation over the standard VAR.

We compare the performance of our models at forecasting the artificial data. At each replication, we simulate $T = 600$ observations of which the last 100 are considered out-of-sample (The size of the in-sample rolling window therefore equals 500). We forecast two series (output growth and inflation growth) at three different horizons ($h = 1, 6$ and 12). The five forecast performance evaluation metrics are computed for each simulated sample and averaged over 100 Monte Carlo replications.\textsuperscript{16}

The results are shown in Figures 18 and 19.\textsuperscript{17} When predicting the output growth, we find that our regularized data-rich techniques consistently produce the best point and sign forecast performance. In particular, models with targeted predictors minimize mean squared and absolute errors, which suggests that not all series are useful. However, pre-selecting variables is not enough given that data-rich model averaging outperforms targeted diffusion indices. Therefore, the combination of regularization and model averaging is needed.\textsuperscript{18} The improvement over the standard autoregressive alternatives, in terms of MSPE, range between 10% and 20% (short and long horizons). In the case of inflation growth, our targeted CSR model performs generally the best for short horizon while univariate iterative alternatives, ARMA and ARI take the lead at horizons 6 and 12. Contrary to aggregate output, considering a large data sets and dimension reduction methods adds only a small improvement.

\textsuperscript{16}The simulation exercise is extremely time consuming: 5 days on a cluster using 20 cores with Matlab R2016. With only 100 Monte Carlo replications, we already have to compute 10 000 forecasts (100 replications multiplied by 100 out-of-sample periods) for each series and each horizon.

\textsuperscript{17}The supplementary material contains tables with complete results. Note that Ruge-Murcia and Onatski (2013) have used another measure of forecast accuracy, the variances of optimal forecast error and the forecast error, since output and inflation are not easily forecastable. We have studied the predictability of these series using the pseudo-$R^2$ and found that the output growth is quite forecastable but not the inflation rate.

\textsuperscript{18}Using this simulation design Stevanovic (2015) has found that several disaggregated series do not have a strong factor structure and that pre-selection improves the estimation of the impulse response functions.
Figure 18: Simulation Evidence: Aggregate Output Growth

Figure 19: Simulation Evidence: Aggregate Inflation Rate
7 Conclusion

This paper compares the performance of six classes of forecasting models on four types of time series in an extensive out-of-sample exercise. The classes of models considered are (i) standard univariate models (Autoregressive Direct, Autoregressive Iterative, Autoregressive Distributed Lag and ARMA(1,1)), (ii) factor-augmented regressions (Diffusion Indices, Targeted Diffusion Indices, Diffusion Indices with dynamic factors and Three-pass Regression Filter), (iii) dynamic factor models (e.g., FAVAR, FAVARMA and DFM), (iv) Data-Rich model averaging (Complete Subset Regression or CSR), (v) Regularized Data-Rich Model Averaging (CSR combined with preselection of variables or with Ridge regularization), and (vi) forecast combinations (naive average, median, trimmed average and inversely proportional average of all forecasts).

The series considered are the Industrial Production growth, the Employment growth, the inflation growth and the SP500 returns. The comparison of the models is based on their pseudo out-of-sample performance along five metrics: the Mean Square Prediction Error, the Mean Absolute Prediction Error, the ratio of correctly predicted signs, the coverage rate of an interval forecast and the p-value of a forecast optimality test à la Mincer-Zarnowitz. The Model Confidence Sets are also considered. For each series, horizon and out-of-sample period, the hyperparameters of our models (number of lags, number of factors, etc.) are re-calibrated using the Bayesian Information Criterion (BIC).

Considering the real series, we find that Forecast Combinations and Regularized Data-Rich Model Averaging generally deliver the best forecasting performance. Data-Rich model averaging techniques are often dominated by their Regularized counterparts while Factor Augmented and Factor Structure-based models often dominate standard univariate models. During recession periods, some factor structure-based and factor-augmented models emerge among the best to predict real series due to their flexibility. Forecast combinations and Regularized Data-Rich models averaging techniques still perform well along each performance metrics.

In case of inflation growth, we find that the ARMA(1,1) model performs incredibly well and generally outperforms most Data-Rich models. We attribute this good performance of the ARMA(1,1) to the fact that inflation anticipations are well anchored so that inflation growth is exogenous with respect to the information set on which the forecasts are based. Forecast combinations are the second-best approaches to predict inflation growth at most horizon, followed by Regularized Data-Rich model averaging. During recessions, the ARMA(1,1) model is often dominated by other alternatives.

Considering the SP500 returns, forecast combinations and Data-Rich model averaging (Regularized and dense) are the generally best forecasting techniques. Factor structure-based and factor augmented models are dominated in general but they often emerge as good benchmarks when the SR metrics is considered. During recessions, forecast combination, Regularized Data-
Rich model averaging and factor augmented models perform well at short horizon but their performance deteriorate at long horizon. Factor structure-based model are slightly dominated at short horizons and are more resilient at longer horizons.

Overall, the family of Regularized Data-Rich model averaging techniques emerges as the most robust of all. Further simulation results based on a large-scale multi-sector Dynamic Stochastic General Equilibrium model show that either regularization alone or model averaging alone is dominated. Indeed, the robustness of the Regularized Data-Rich model averaging techniques is due to the fact that they combine the two features.

Finally, we examine the stability the forecasting equations and their performance over time. The results suggest a lot of time instability in the forecast accuracy as well as in the structure of the optimal forecasting equations.

References


