Kotchoni, Rachidi; Leroux, Maxime; Stevanovic, Dalibor

Working Paper

Macroeconomic Forecast Accuracy in a Data-Rich Environment

Document de travail, No. 2017-02

Provided in Cooperation with:
Department of Economics, School of Management Sciences (ESG UQAM), University of Quebec in Montreal

Suggested Citation: Kotchoni, Rachidi; Leroux, Maxime; Stevanovic, Dalibor (2017) : Macroeconomic Forecast Accuracy in a Data-Rich Environment, Document de travail, No. 2017-02, Université du Québec à Montréal, École des sciences de la gestion (ESG UQAM), Département des sciences économiques, Montréal

This Version is available at:
https://hdl.handle.net/10419/234747

ESG RECHERCHE

DOCUMENT DE TRAVAIL / WORKING PAPER

No. 2017-02

Macroeconomic Forecast Accuracy in a Data-Rich Environment

Maxime Leroux, Rachidi Kotchoni, *et* Dalibor Stevanovic

Septembre 2017

ESG UQÀM

Département des sciences économiques
École des sciences de la gestion
Université du Québec à Montréal

Macroeconomic Forecast Accuracy in a Data-Rich Environment

Maxime Leroux, Université du Québec à Montréal, Canada
Rachidi Kotchoni, Economix-CNRS, Université Paris Nanterre, France
Dalibor Stevanovic, Université du Québec à Montréal, Canada

Document de travail No. 2017-02

Septembre 2017

Département des Sciences Économiques
Université du Québec à Montréal
Case postale 8888,
Succ. Centre-Ville
Montréal, (Québec), H3C 3P8, Canada
Courriel : brisson.lorraine@uqam.ca
Site web : http://economie.esg.uqam.ca

Macroeconomic Forecast Accuracy in a Data-Rich Environment

De l'information supplémentaire peut être accédée à / supporting information can be found at:
http://www.stevanovic.uqam.ca/LKS_ForecastingDataRich_SuppMaterial.pdf

Les codes Matlab sont disponibles à / Matlab codes are available at:
http://www.stevanovic.uqam.ca

Macroeconomic forecast accuracy in a data-rich environment*

Maxime Leroux† Rachidi Kotchoni‡ Dalibor Stevanovic §

This version: September 15, 2017

Abstract

We compare the performance of six classes of models at forecasting different types of economic series 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 cross-sectional dispersion of out-of-sample point forecasts is a good predictor of macroeconomic uncertainty. *(v)* 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, Domenico Giannone, Serena Ng, Frank Schorfheide and participants of the Penn Big Data Conference 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

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, Lenza & Primiceri 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 & Ng 2005), (Kim & Swanson

[1] Bayesian techniques developed in recent years to handle larger than usual VAR models can be viewed as an effort toward this objective. See (Banbura, Giannone & Reichlin 2010), (Koop 2013), (Carriero, Clark & Marcellino 2015) and (Giannone, Lenza & Primiceri 2015), among others.

[2] These 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.

2014), (Cheng & Hansen 2015), (Carrasco & Rossi 2016) and (Groen & Kapetanios 2016).

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 & Watson 2002b), the Targeted DI of (Bai & Ng 2008), the DI with dynamic factors of (Forni, Hallin, Lippi & Reichlin 2005) and the Three-pass Regression Filter (3PRF) of (Kelly & 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 & Ng 2005), the Factor-Augmented VARMA (FAVARMA) of (Dufour & 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, Gargano & Timmermann 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).[3] 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 & Watson 2004).[4]

The monthly macroeconomic data employed for this study comes from (McCracken & 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 the first 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 future devel-

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

[4]For the sake of completeness, the simple random walk (RW) and the random walk with drift (RWD) are considered as well when relevant.

opment of forecasting models.[5] 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 nonlinearity in the dynamics of the target. 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. 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 generalization performance.[6] 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.

Over the full out-of-sample period, 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 RW model delivered 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. The cross-sectional dispersion of our point forecasts changes over time as well. Indeed, it is significantly correlated with some

[5]The data used in this paper are publicly available. Our Matlab codes are available upon a simple request.
[6]References on inflation forecasting include (Stock & Watson 2007) and (Faust & Wright 2013).

macroeconomic and financial uncertainty measures used in the literature.

In order to verify the robustness of our empirical results, we compare the performance of the forecasting models using data that are simulated from a data generating process (DGP) implied by a large-scale Dynamic Stochastic General Equilibrium (DSGE) model proposed by (Ruge-Murcia & 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.

The remainder of the paper is organized as follows. Section 2 presents the standard time series models considered in this paper. Section 3 presents the Data-Rich models and Section 4 presents the Forecasts Combinations techniques. Section 5 presents the data, the design of the pseudo out-of-sample exercise and the forecast evaluation metrics. Section 6 reports the main empirical results. Section 7 presents further simulation results while Section 8 analyzes the cross-sectional dispersion of the forecasts. Section 9 analyzes the stability of the forecast accuracy of the models over time and Section 10 concludes. Additional results are available in supplementary materials.

2 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:

$$y_{t+h}^{(h)} = (freq/h) \sum_{k=1}^{h} y_{t+k}, \tag{1}$$

where $y_t \equiv \ln Y_t$ and $freq$ depends on the frequency of the data (400 if Y_t is quarterly, 1200 if Y_t is monthly, etc.).

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 & Watson 2002b) and (McCracken & Ng 2015). We shall therefore define $y_{t+h}^{(h)}$ as:

$$y_{t+h}^{(h)} = (freq/h) \sum_{k=1}^{h} y_{t+k} = (freq/h) \ln(Y_{t+h}/Y_t), \tag{2}$$

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 $y_{t+h}^{(h)}$ as:

$$y_{t+h}^{(h)} = (freq/h) \sum_{k=1}^{h} y_{t+k} = (freq/h) \left[\ln(Y_{t+h}/Y_{t+h-1}) - \ln(Y_t/Y_{t-1}) \right], \tag{3}$$

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

Indeed, $y_{t+h}^{(h)}$ 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 univariate and multivariate 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:

$$y_{t+h}^{(h)} = \alpha^{(h)} + \sum_{l=1}^{L} \rho_l^{(h)} y_{t-l+1} + e_{t+h}, \quad t = 1, \ldots, T, \tag{4}$$

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

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

where $\hat{\alpha}^{(h)}$ and $\hat{\rho}^{(h)}$ are OLS estimators of $\alpha^{(h)}$ and $\rho^{(h)}$. 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:

$$y_{t+1} = \alpha + \sum_{l=1}^{L} \rho_l y_{t+1-l} + e_{t+1}, \quad t = 1, \ldots, T. \tag{5}$$

where $L \geq 1$. This model is termed *autoregressive iterative* (ARI) because $\widehat{y}_{T+h|T}^h$ must be deduced from recursive calculations of $\widehat{y}_{T+1|T}$, $\widehat{y}_{T+2|T}$,...,$\widehat{y}_{T+h|T}$. We have:

$$\hat{y}_{T+k|T} = \hat{\alpha} + \sum_{l=1}^{L} \hat{\rho}_l \widehat{y}_{T+k-l|T}, \quad k = 1, ..., h,$$

with the convention $\widehat{y}_{t|T} \equiv y_t$ for all $t \leq T$ and:

$$\widehat{y}_{T+h|T}^h = (freq/h) \sum_{k=1}^{h} \hat{y}_{T+k|T}. \tag{6}$$

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.[7]

ARMA(1,1) (Dufour & 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.[8] The following representation is therefore considered and estimated by maximum likelihood:

$$y_{t+1} = \alpha + \rho y_t + \theta e_t + e_{t+1}. \tag{7}$$

After estimation, the residuals $\hat{e}_T$ of the ARMA(1,1) model are generated in-sample using the recursion starting from the initial value $\widehat{e}_1 = 0$:

$$\widehat{e}_{t+1} = y_{t+1} - \widehat{\alpha} - \widehat{\rho} y_t - \widehat{\theta}\widehat{e}_t, \quad t = 1, ..., T.$$

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

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

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

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:

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

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

[7] If the true DGP of y_t is an AR(L), both the direct and iterative approaches should produce the same predictions for any horizon asymptotically as the sample size goes to infinity. However, none of the two specifications strictly dominates in finite samples. 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, Stock & Watson 2006) compare the forecasting performance of direct and iterative models for hundreds of time series. They conclude that the direct approach provides slightly better results but does not dominate uniformly across time and series.

[8] The ARMA(1,1) model has been used extensively in the empirical finance literature to forecast the realized volatility, but has been considered much less for the prediction of macroeconomic series.

3 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 & Watson 2002) and (Stock & 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 & Stevanovic 2013) and (Stock & Watson 2005), we assume that:

$$\begin{aligned} X_t &= \lambda(L) f_t + u_t \,, && (9)\\ u_t &= \delta(L) u_{t-1} + \nu_t \,, && (10)\\ f_t &= \gamma(L) f_{t-1} + \theta(L)\eta_t \,, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, && (11) \end{aligned}$$

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

We have:

$$\begin{aligned} \underset{(N\times q)}{\lambda(L)} &= \sum_{k=0}^{p_\lambda - 1} \lambda_k L^k; \qquad \underset{(N\times N)}{\delta(L)} = \sum_{k=0}^{p_\delta - 1} \delta_k L^k, \\ \underset{(q\times q)}{\gamma(L)} &= \sum_{k=0}^{p_\gamma - 1} \gamma_k L^k; \qquad \underset{q\times q}{\theta(L)} = I_q - \sum_{k=1}^{p_\theta} \theta_k L^{q_f} \end{aligned}$$

with p_λ, p_δ, p_γ, $p_\theta \geq 1$ are the highest degrees of polynomials in each matrix. Indeed, the matrices of coefficients λ_k, δ_k, γ_k and θ_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:

$$\begin{aligned} X_{it} &= \sum_{k=0}^{p_\lambda - 1} \lambda_{k,i} f_{t-k} + u_{i,t} \equiv \lambda^{(i)}(L) f_t + u_{it}, && (12)\\ u_{it} &= \sum_{k=0}^{p_\delta - 1} \delta_{k,i} u_{i,t-1-k} + \nu_{it} \equiv \delta^{(i)}(L)\, u_{i,t-1} + \nu_{it}, && (13) \end{aligned}$$

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

The exact DFM is obtained if the following assumption is satisfied:

$$E(u_{it}u_{js}) = 0\,,\ \forall i,j,t,s, \quad i \neq j.$$

The approximate DFM is obtained by allowing for some limited cross-section correlations among the idiosyncratic components.[9] We assume the idiosyncratic errors ν_{it} are 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:

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

where

$$\underset{(N\times qp_\lambda)}{\Lambda} = \begin{bmatrix} \lambda_0 & \lambda_1 & \ldots & \lambda_{p_\lambda-1} \end{bmatrix}$$

$$\underset{(qp_\lambda\times qp_\lambda)}{\Gamma(L)} = \begin{bmatrix} \gamma_0 & \gamma_1 & \ldots & \gamma_{p_\gamma-1} \\ 0 & I & 0 & \ldots \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \ldots & 0 & I \end{bmatrix};\quad \underset{(qp_\lambda\times q)}{\Theta(L)} = \begin{bmatrix} \theta(L) \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Equations (14)-(16) define the FAVARMA model proposed in (Dufour & 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 & 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 & Ng 2005) and to study the impact of monetary policy shocks in (Bernanke, Boivin & Eliasz 2005).

In practice, q and p_λ 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:

$$F_t = \Phi(L)F_{t-1} + \Theta(L)\eta_t, \tag{17}$$

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

[9] Intuitively, only a small number of largest eigenvalues of the covariance matrix of the common component, $\tilde{\lambda}_i(L)f_t$, may diverge when the number of series tends to infinity, while the remaining eigenvalues as well as the eigenvalues of the covariance matrix of specific components are bounded. See technical details in (Stock & Watson 2005) and (Bai & Ng 2008).

of p_ϕ and p_θ can be selected by BIC.

3.1 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 & Watson 2002b):

$$y_{t+h}^{(h)} = \alpha^{(h)} + \sum_{l=1}^{p_y^h} \rho_l^{(h)} y_{t-l+1} + \sum_{l=1}^{p_f^h} F_{t-l+1}\beta_l^{(h)} + e_{t+h}, \quad t = 1, \ldots, T \tag{18}$$

$$X_t = \Lambda F_t + u_t \tag{19}$$

where F_t are $K^{(h)}$ *consecutive static* factors and the superscript h stands for the value of K when forecasting h periods ahead. The optimal values of p_y^h, p_f^h and $K^{(h)}$ are simultaneously selected by BIC. The h-step ahead forecast is obtained as:

$$\hat{y}_{T+h|T}^h = \hat{\alpha}^{(h)} + \sum_{l=1}^{p_y^h} \widehat{\rho}_l^{(h)} y_{T-l+1} + \sum_{l=1}^{p_f^h} F_{T-l+1}\widehat{\beta}_l^{(h)}.$$

The feasible ARDI model is obtained after estimating F_t as the first $K^{(h)}$ principal components of X_t. See (Stock & Watson 2002) 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:

$$\begin{aligned} \hat{y}_{T+h|T}^{h} &= \hat{\alpha}^{(h)} + \sum_{l=1}^{p_y^h} \widehat{\rho}_l^{(h)} y_{T-l+1} + \sum_{i \in K^*} \hat{F}_{i,T} \beta_i^{(h)} \\ K^* &= \{i \in 1, \ldots, K \mid t_i > t_c\}. \end{aligned} \tag{20}$$

where $K^* \in K$ refers to elements of F_t corresponding to coefficients β_i^h 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 & 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 $y_{t+h}^{(h)}$. The ARDIT model of (Bai & 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 $y_{t+h}^{(h)}$ and next predict the factors using this subset. (Bai & Ng 2008) propose two ways to construct the subset X_t^*:

- Hard threshold (OLS): **ARDIT-hard**

$$y_{t+h}^{(h)} = \alpha^{(h)} + \sum_{j=0}^{3} \rho_j^{(h)} y_{t-j} + \beta_i^{(h)} X_{i,t} + \epsilon_t \tag{21}$$

$$X_t^* = \{X_i \in X_t \mid t_{Xi} > t_c\} \tag{22}$$

- Soft threshold (LASSO): **ARDIT-soft**

$$\hat{\beta}^{lasso} = arg\, min_\beta \left[RSS + \lambda \sum_{i=1}^{N} |\beta_i| \right] \tag{23}$$

$$X_t^* = \{X_i \in X_t \mid \beta_i^{lasso} \neq 0\} \tag{24}$$

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 & 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.[10]

Three-Pass Regression Filter (3PRF) (Kelly & 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 $i = 1, \ldots, N$

$$X_{i,t} = \phi_{0,i} + Z_t'\phi_i + \varepsilon_{i,t}$$

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

$$\hat{y}_{T+h|T}^{(h)} = \hat{\beta}_0 + \hat{\beta}\hat{f}_T$$

We follow (Kelly & 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.

3.2 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:

$$X_t = \Lambda F_t + u_t \tag{25}$$

$$u_t = \delta(L)u_{t-1} + v_t \tag{26}$$

$$F_t = \Phi F_{t-1} + \eta_t. \tag{27}$$

[10] As in (Bai & Ng 2008) we target 30 series. It is possible to optimally select the number of retained series, but the procedure is very long and (Bai & Ng 2008) did not find significant improvements.

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 & Ng 2002) IC_{p2} criterion. After estimation, one forecasts the factors using (27) upon assuming stationarity. The idiosyncratic component is predicted using (26) and then $\widehat{F}_t$ and $\widehat{u}_t$ are combined into (25) to obtain a prediction of X_t. (Boivin & Ng 2005) compare the direct and iterative approaches:

- Iterative

$$\begin{aligned}\hat{F}_{T+h|T} &= \widehat{\Phi}\hat{F}_{T+h-1|T}\\ \hat{u}_{T+h|T} &= \hat{\delta}(L)\hat{u}_{T+h-1|T}\\ \hat{X}_{T+h|T} &= \hat{\Lambda}\hat{F}_{T+h|T} + \hat{u}_{T+h|T}\end{aligned}$$

- Direct

$$\begin{aligned}\hat{F}^{(h)}_{T+h|T} &= \widehat{\Phi}^{(h)}\hat{F}_T\\ \hat{u}^{(h)}_{T+h|T} &= \hat{\delta}(L)^{(h)}\hat{u}_T\\ \hat{X}^{(h)}_{T+h|T} &= \hat{\Lambda}\hat{F}^{(h)}_{T+h|T} + \hat{u}^{(h)}_{T+h|T}\end{aligned}$$

The forecast of interest, $\hat{y}^{(h)}_{T+h|T}$, is then extracted from $\hat{X}_{T+h|T}$ or $\hat{X}^{(h)}_{T+h|T}$. 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 & 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 estimable 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:

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

with $\hat{\eta}_{T+h-k|T} = 0$ if $h - k > 0$. The forecast $\hat{y}^h_{T+h|T}$ 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.[11] The forecasting formula for the idiosyncratic component remains the same. The forecast of X_t is constructed as follows:

$$\begin{aligned} \hat{u}_{T+h|T} &= \hat{\delta}(L)\hat{u}_{T+h-1|T} \\ \hat{X}_{T+h|T} &= \widehat{\lambda}(L)\widehat{f}_{T+h|T} + \hat{u}_{T+h|T} \end{aligned}$$

and $\hat{y}^{(h)}_{T+h|T}$ is extracted from $\hat{X}_{T+h|T}$. The number of underlying dynamic factors f_t is selected by (Hallin & 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.

3.3 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 $y^{(h)}_{T+h|T}$ 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:

$$\hat{y}^{(h)}_{T+h|T,m} = \hat{c} + \hat{\rho} y_t + \hat{\beta} X_{t,m} \tag{28}$$

$$\hat{y}^{(h)}_{T+h|T} = \frac{\sum_{m=1}^{M} \hat{y}_{T+h|T,m}}{M} \tag{29}$$

where $X_{t,m}$ contains L series for each model $m = 1, \ldots, M$.[12] This method can be computationally demanding when the number of predictors in X_t is large.

3.4 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).

[11] See (Boivin & Ng 2005) for discussion. It is the 'DN' specification in their paper.

[12] 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).

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

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

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

As suggested by (Bai & 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 Targeting 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*.

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

Step 2 CSR on X_t^*

Following (Bai & 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. 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 different specifications of Ridge CSR: one based on 10 regressors and another based on 20 regressors.

4 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:

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

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 $\left(y_{t+h|t}^{(h,1)} \leq y_{t+h|t}^{(h,2)} \ldots \leq y_{t+h|t}^{(h,M)}\right)$. Then trim a proportion λ of forecasts from both sides:

$$y_{t+h|t}^{(h,trim)} = \frac{1}{M(1-2\lambda)} \sum_{i=\lceil \lambda M \rceil}^{\lfloor (1-\lambda)M \rfloor} y_{t+h|t}^{(h,i)}$$

where $\lceil \lambda M \rceil$ is the integer immediately larger than λM 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 & Pauly 1987). Here, we follow (Stock & 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 ρ 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:

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

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

5 Empirical Evaluation of the Forecasting Models

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

5.1 Data

We use historical data to evaluate and compare the performance of all the forecasting models described previously.[13] 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 & 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 & Ng 2015) database, but this can be accommodated when a rolling window is used. Indeed, a series that is not available at the starting date will eventually appear in the informational set as the window moves forward.[14]

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 & Ng 2015) and (Stock & Watson 2002b) and assume that price indexes are all $I(2)$ while interest and unemployment rates are $I(1)$. [15]

[13] In principle, a real-time forecasting exercise could be preferable but not all variables are available in real-time vintages. Hence, we choose to evaluate the models with the most recent releases and not consider their performance in the presence of revisions.

[14] However, this is a problem when conducting a structural FAVAR analysis as in (Bernanke et al. 2005). Another source of unbalanced panels is mixing frequencies. (Stock & Watson 2002b) construct a monthly data set using monthly and quarterly series. They transform the quarterly series into monthly indicators using an expectation-maximization (EM) procedure that also works to fill the holes of unobserved monthly data points. This EM technique has also been used in (Boivin, Giannoni & Stevanović 2013) when estimating the effects of credit shocks.

[15] (Bernanke et al. 2005) keep inflation, interest and unemployment rates in levels in X_t. Choosing (SW) or (BBE) transformations has important effects on correlation patterns in X_t. Under (BBE), the group of interest rates is highly correlated as well as the inflation and unemployment rates. Hence, the principal components will tend to exploit these clusters such that the initial factors will be related to those groups of series. As pointed out by (Boivin & Ng 2006), the presence of these clusters may alter the estimation of *common* factors. Under (SW), these correlation clusters are less important. Recently, procedures have been proposed to deal directly with the unit root instead of differentiating the data, see (Banerjee, Marcellino & Masten 2014) and (Barigozzi, Lippi & Luciani 2016).

5.2 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.[16] 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, where h is the forecasting horizon.[17]

5.3 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 the SP500 are treated as I(1) while the logarithm of the CPI is assumed to be I(2), as in (Stock & Watson 2002b) and (McCracken & Ng 2015).

5.4 Forecast Evaluation Metrics

The forecasting models will be compared using five metrics. Two of these metrics 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.

5.4.1 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.[18]

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

[16] Further in the paper we compare the forecast accuracy of rolling versus expanding (or recursive) windows.

[17] (Inoue, Kilian & Rossi 2016) propose to optimally select the window size in the presence of structural breaks.

[18] The supplementary reports the pseudo-R^2, a related measure of predictability from (Galbraith 2003).

5.4.2 Interval Forecast Evaluation

Ideally, we would like to have predictive densities in addition to point forecasts. However, density forecasts are generally harder to produce than point forecasts. We elude this problem by resorting to simplifying parametric assumptions. Let $\widehat{y}_{t+h|t}$ be the point forecast and σ_h^2 the associated variance. Assuming normality for the forecasting errors leads to:

$$y_{t+h} \sim N(\widehat{y}_{t+h|t}, \sigma_h^2). \tag{30}$$

Hence, an interval forecast can be deduced as $\widehat{y}_{t+h|t} \pm c \times \sigma_h$ 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, $\widehat{y}_{t+h|t}$ and σ_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 σ_h^2. Subsequently, this estimate is updated recursively.

5.4.3 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 & Timmermann 1995). This possibility can be assessed by means of the (Pesaran & Timmermann 1992) sign forecast test. The test statistic is given by:

$$S_n = \frac{\hat{p} - \hat{p}^*}{\sqrt{Var(\hat{p}) - Va(\hat{p}^*)}},$$

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 $S_n \longrightarrow \mathrm{N}(0,1)$.[19]

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.

[19]Let q denote the proportion of positive realizations in the actual data and $\widehat{q}$ the proportion of positive forecasts. Under H_0, the estimated theoretical number of correctly signed forecast is $\hat{p}^* = q\widehat{q} + (1-q)(1-\widehat{q})$.

5.4.4 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:

$$y_{t+h} = \beta_0 + \beta_1 \hat{y}_{t+h|t} + u_{t+h}, \tag{31}$$

where $\widehat{y}_{t+h|t}$ is an out-of-sample forecast of y_{t+h}. If $\widehat{y}_{t+h|t}$ 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.[20] Below, the results of the optimality tests are presented separately for each horizon while the joint tests are deferred to the supplementary material.

6 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.[21] 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). Finally, 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

[20] An approach to extend this method to a joint test for all horizons consists of using Bonferroni's bounds. First, one collects the p-values of the chi-squared tests performed after each Mincer-Zarnowitz regression. Next, one rejects the optimality of forecasts if the minimum p-value across all horizons is less than α divided by the number of tests. This approach is known to be conservative. See (Patton & Timmermann 2012) who study this type of optimality testing.

[21] This rescaling is needed to uniformise the presentation.

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.[22]

6.1 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.

[22]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 & Mariano 1995) (DM) test.

6.2 Employment Growth

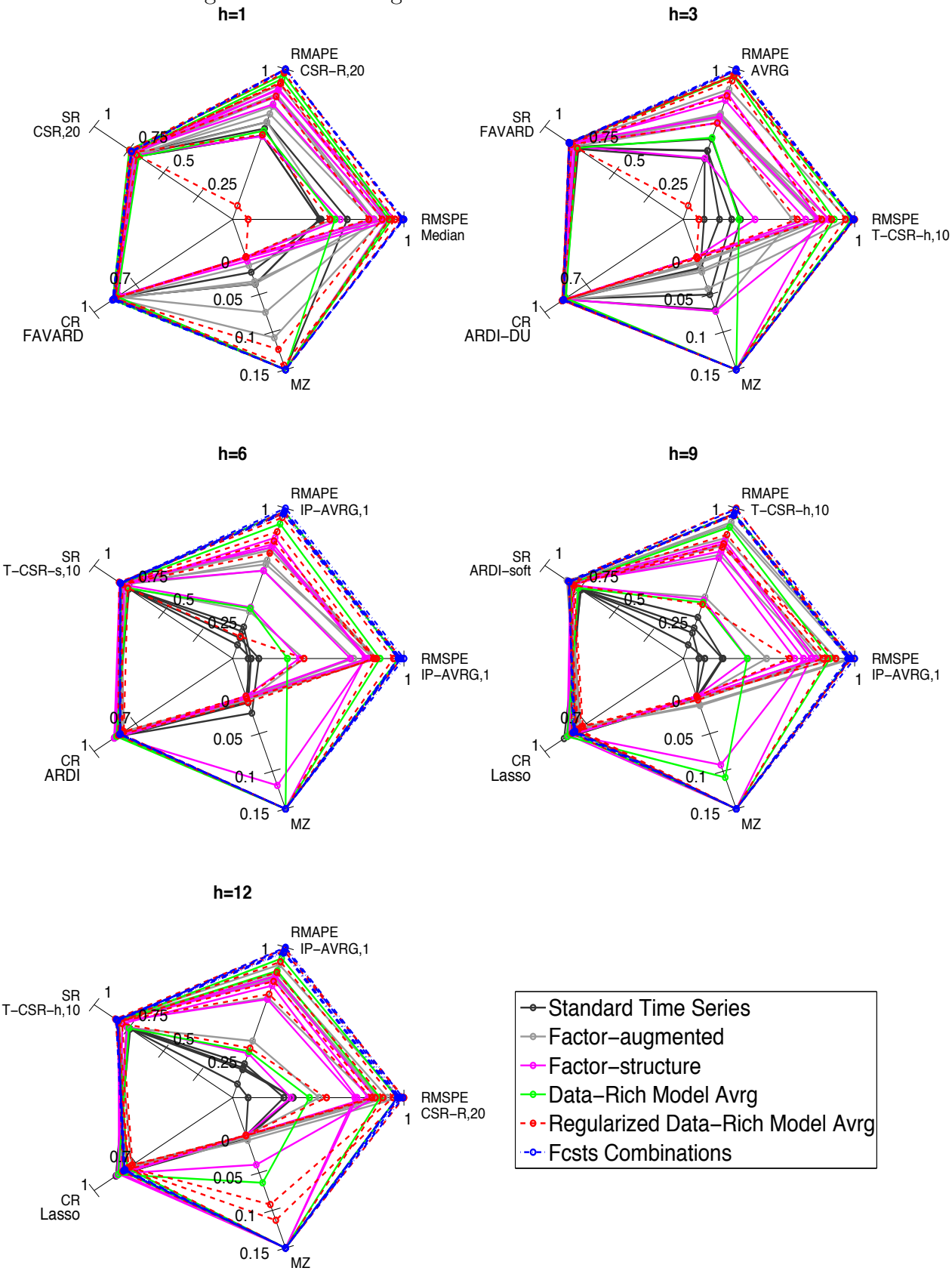
We now examine the empirical results for Employment Growth. Figure 3 shows the results for the Full Out-of-sample period while Figure 4 focuses on data points that belong to NBER recessions. The results are quite similar to what is obtained for industrial production growth.

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 RMSPE 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.

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 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

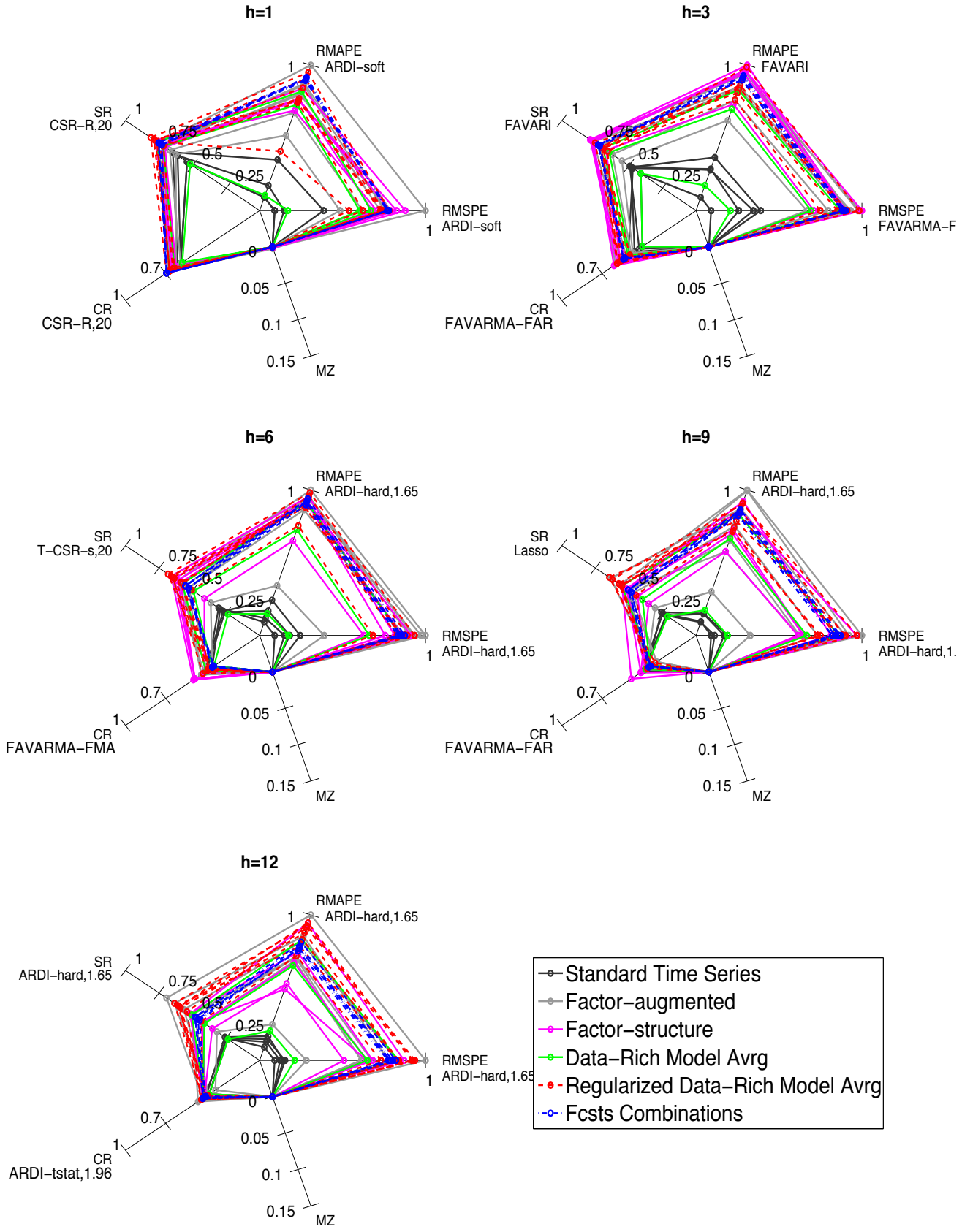


Figure 3: Forecasting Employment: Full OOS

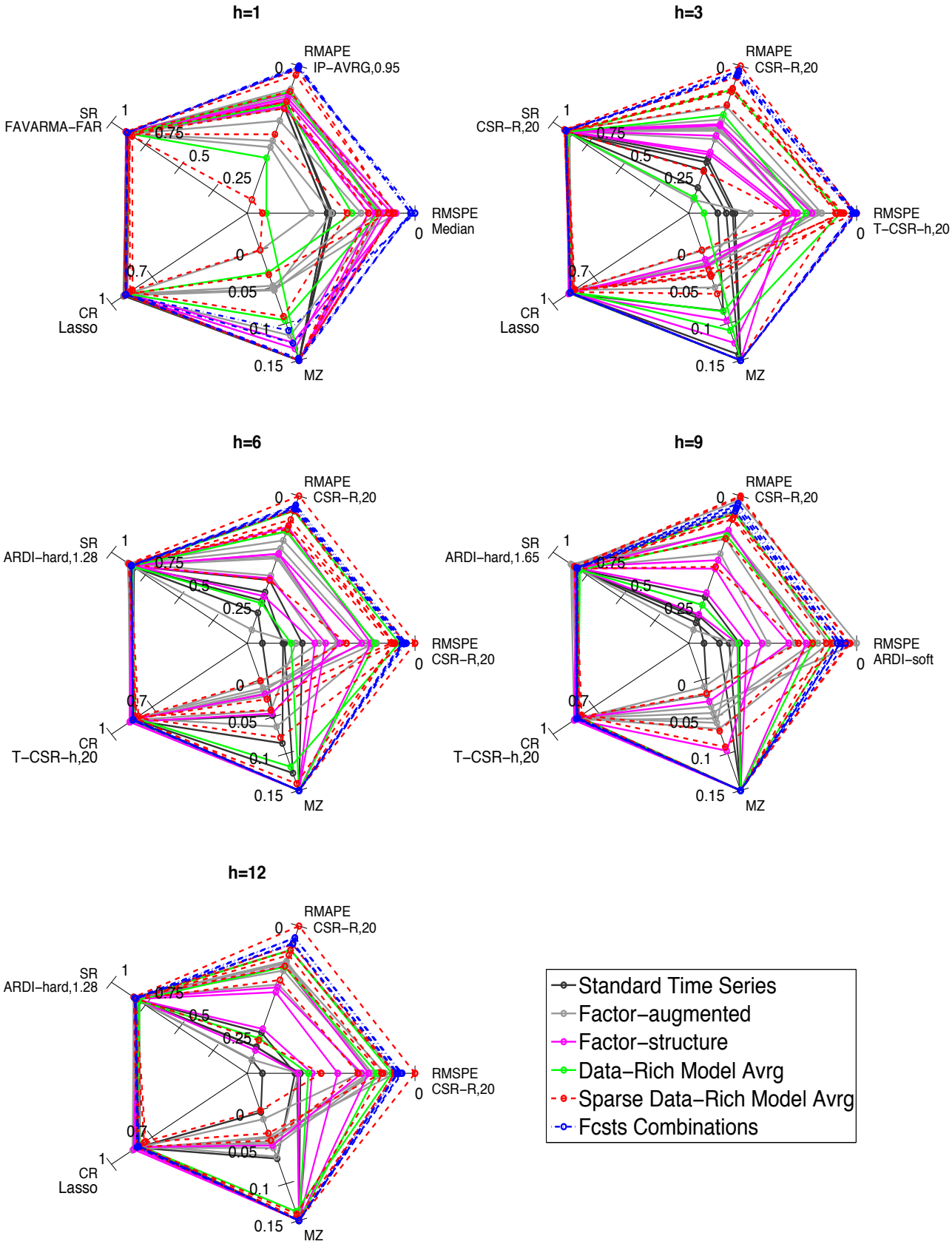
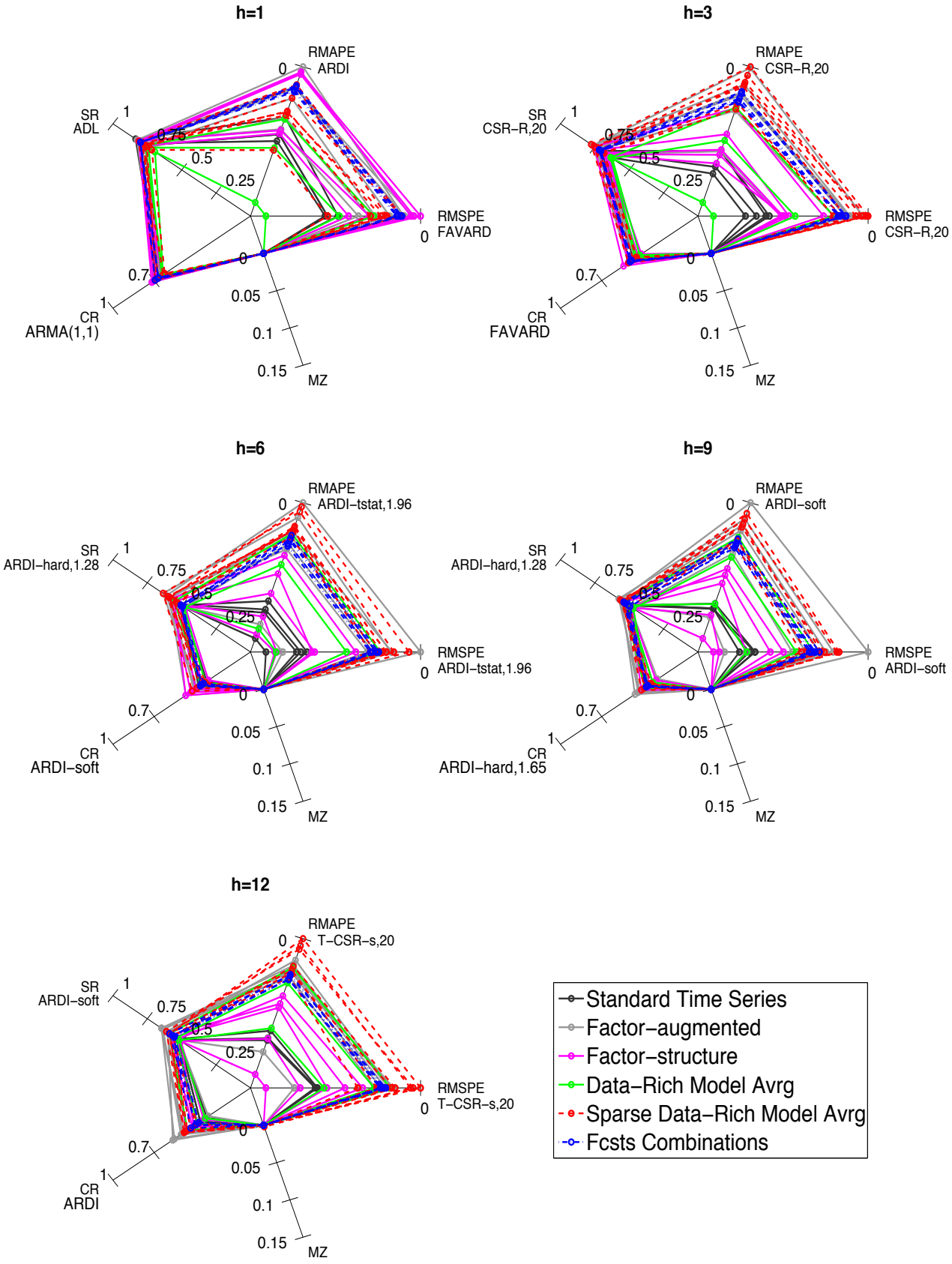


Figure 4: Forecasting Employment: NBER Recession



6.3 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 5 shows the results for the entire out-of-sample period while Figure 6 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 $h = 3$.[23] 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 generalization performance for this series.[24] 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 looses its predictive power and data-rich models become favored.

Overall, 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 situations where the main object of interest is the direction of inflation change, some Data-Rich models should be preferred to the ARMA(1,1) model.

[23](Stock & Watson 2007) suggest that the MA part of the time-varying integrated moving average process of inflation rate has increased from 1984. (Ng & Perron 1996) and (Ng & Perron 2001) also document the importance of the MA component for the U.S. inflation. (Foroni, Marcellino & Stevanovic 2017) found that the MA part improves the forecasting power of mixed-frequency models when predicting the U.S. inflation.

[24]Indeed, a lack of parsimony can cause a model to have good in-sample fit but low out-of-sample performance.

Figure 5: Forecasting CPI Inflation: Full OOS

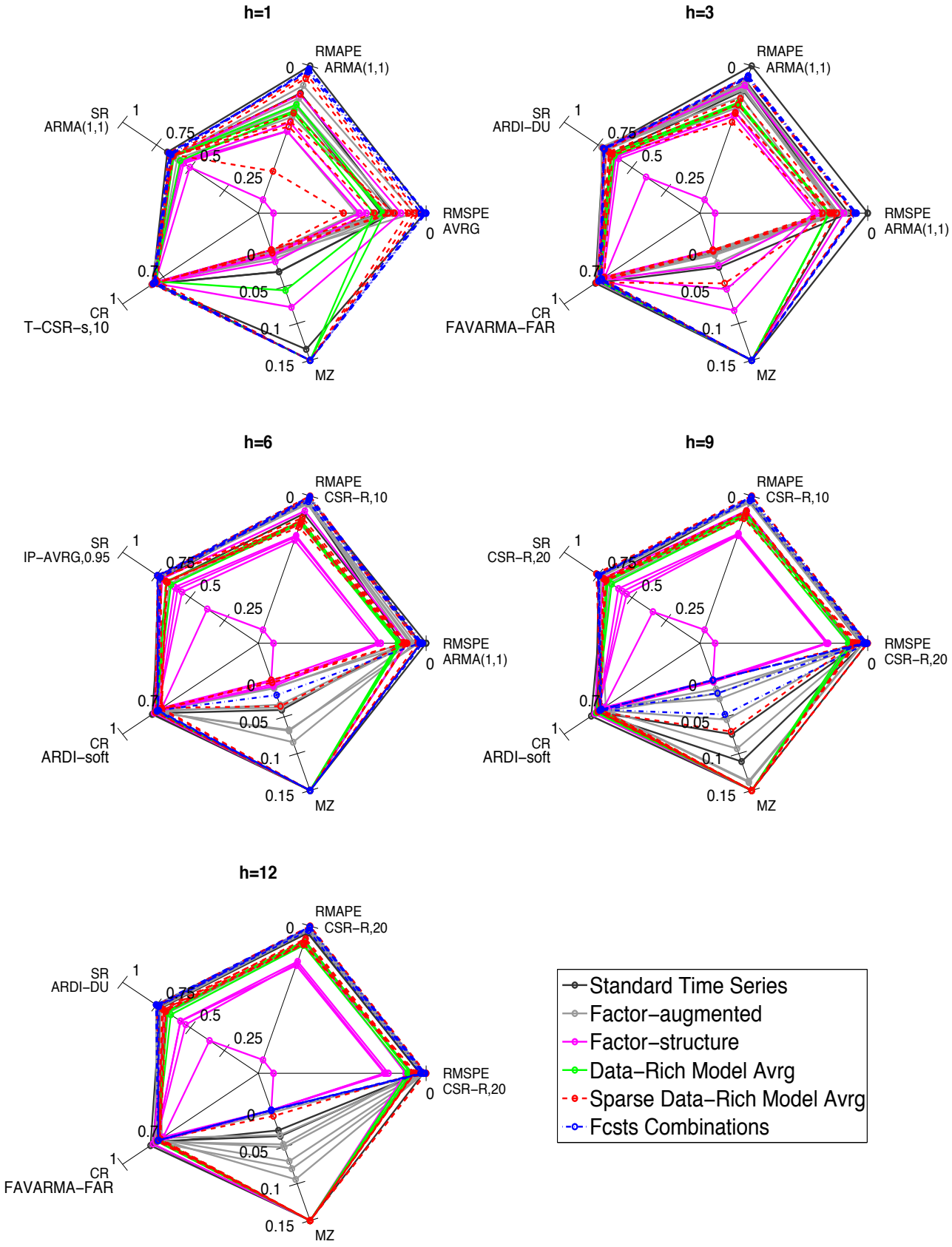
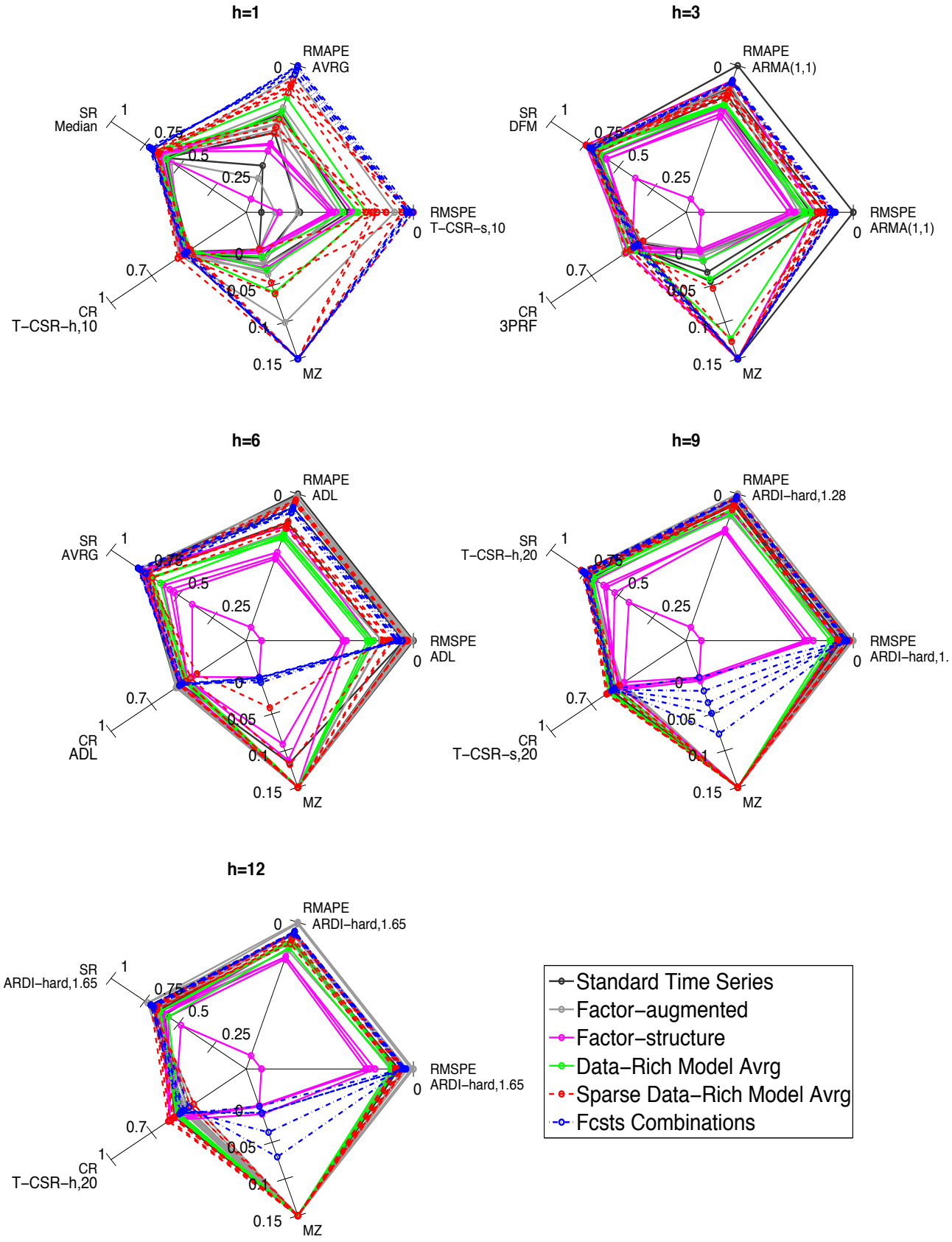


Figure 6: Forecasting CPI Inflation: NBER Recession



6.4 Stock Market Index

We now examine the empirical results for the SP500 returns. Figure 7 shows the results for the entire out-of-sample period while Figure 8 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 (at least at 15% significance level). 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.[25]

[25](Rapach & Zhou 2010) also found that stock market returns can be predictable during recessions.

Figure 7: Forecasting SP500: Full OOS

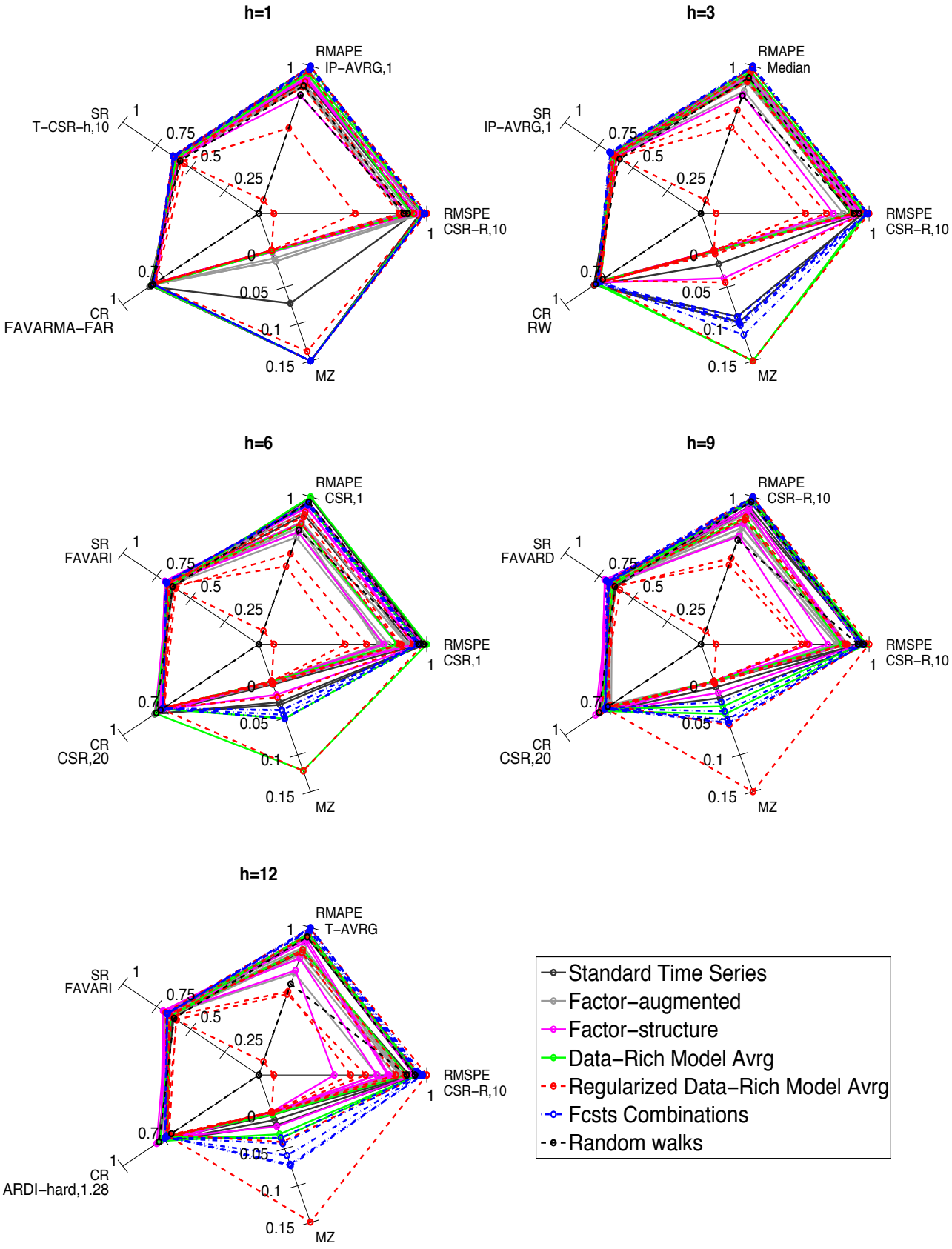
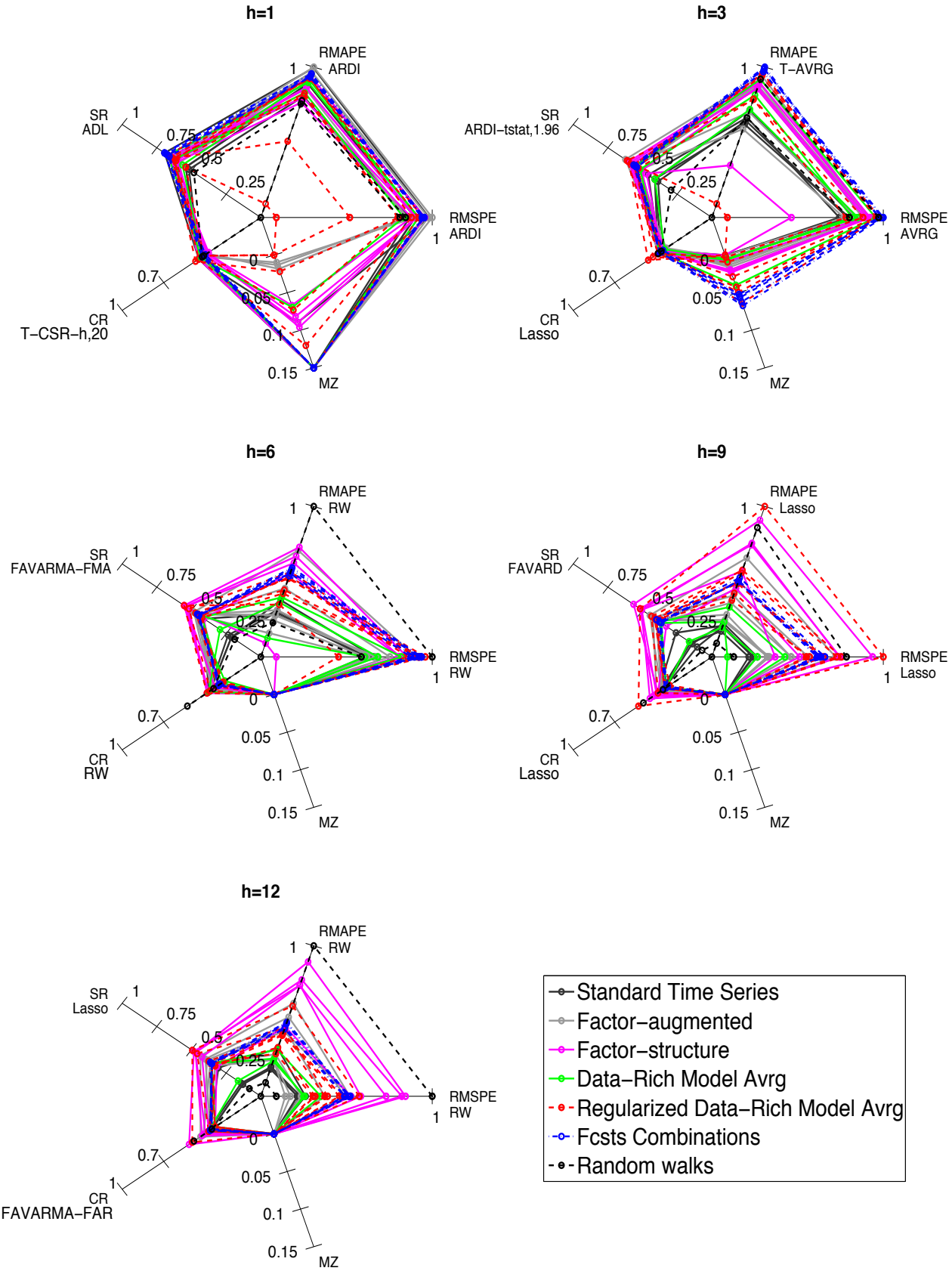


Figure 8: Forecasting SP500: NBER Recession



7 Simulation Evidence

In order to verify the robustness of our empirical results, we simulated artificial data from a model that is not loaded in favor of a particular forecasting technology. We use a data generating process (DGP) that is deduced from a large multi-sector Dynamic Stochastic General Equilibrium (DSGE) model proposed by (Ruge-Murcia & 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. Similar studies in the literature have often considered linear static models or DGPs that involved dynamic factors with several degrees of serial and cross-sectional dependence.[26] Data are simulated from a linear state-space representation with three pervasive common dynamic shocks and 30 sectoral productivity shocks. (Ruge-Murcia & 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.[27]

The results are shown in Figures 9 and 10.[28] 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.[29] 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.

[26] See simulation designs in (Mao Takongmo & Stevanovic 2015) for example.

[27] 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.

[28] The supplementary material contains tables with complete results. Note that (Ruge-Murcia & 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.

[29] 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 9: Simulation Evidence: Aggregate Output Growth

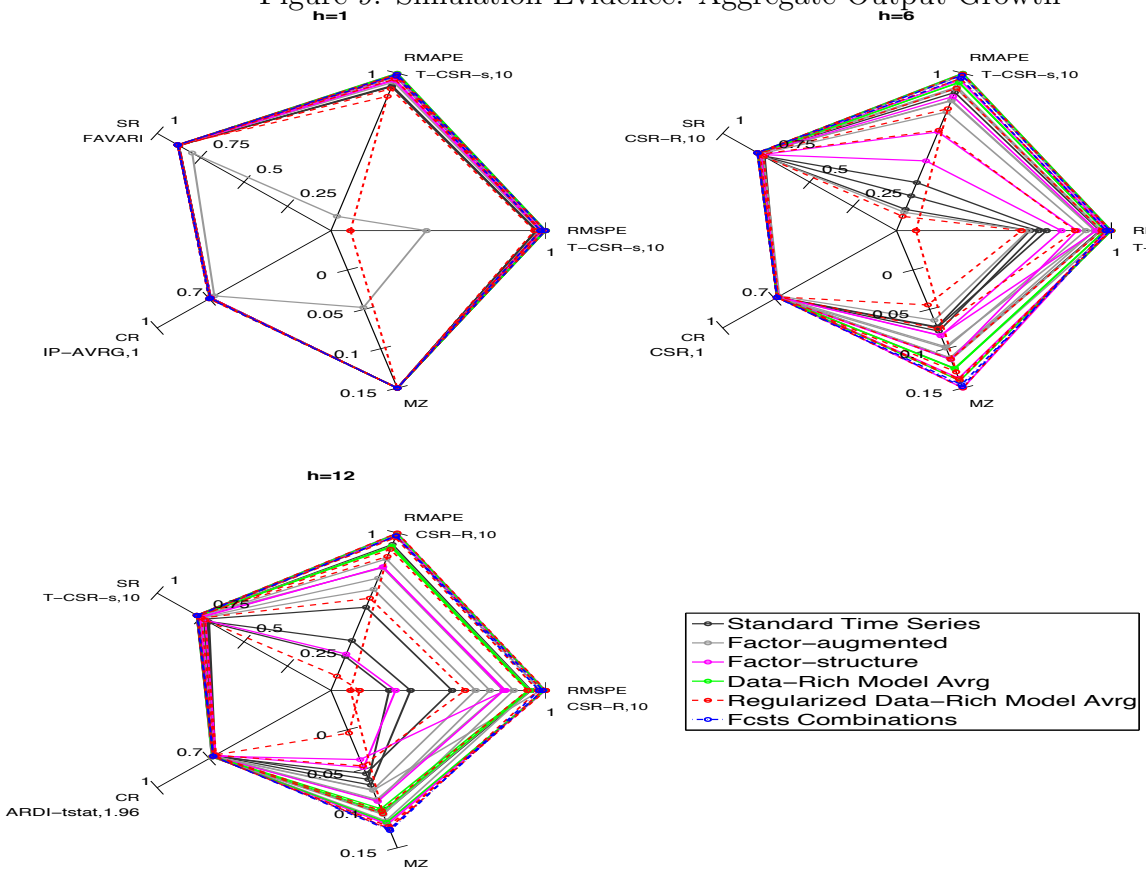
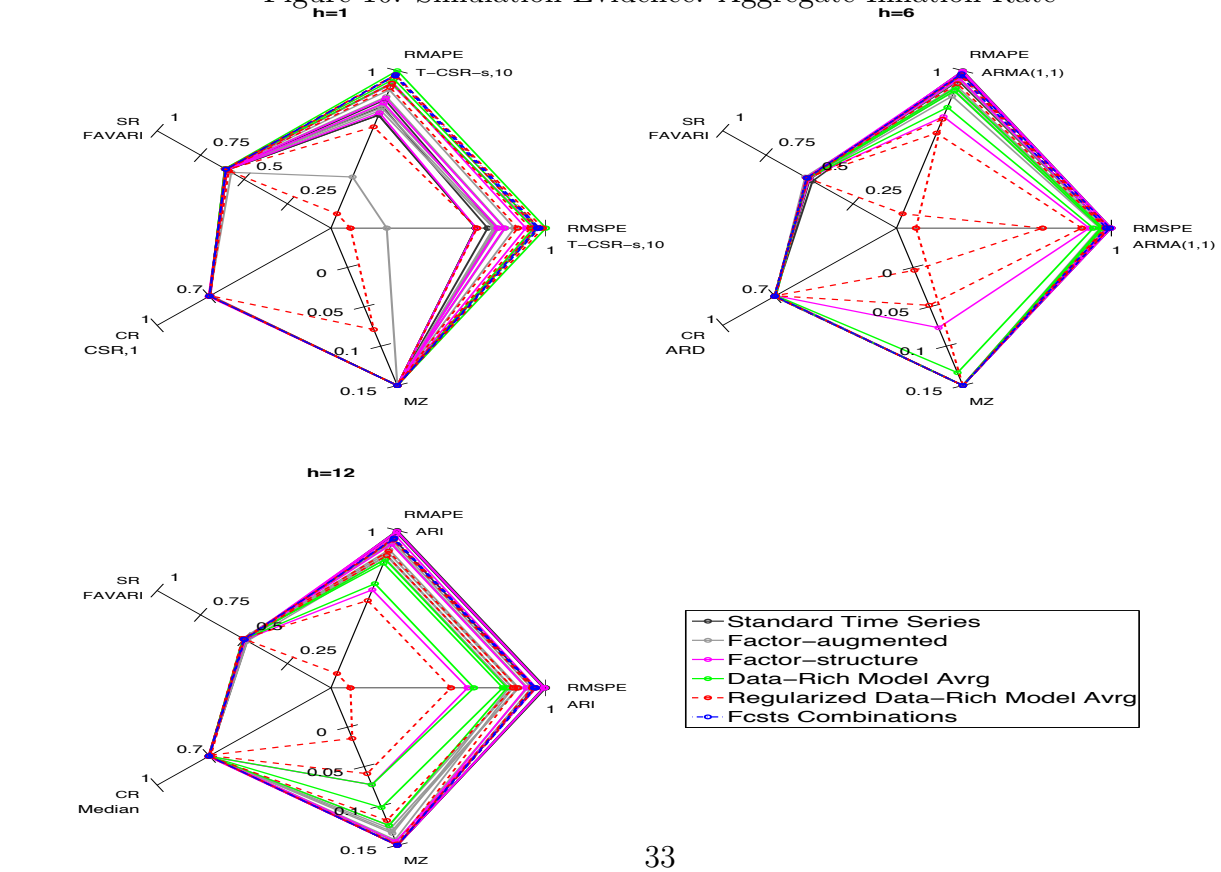


Figure 10: Simulation Evidence: Aggregate Inflation Rate



8 Forecasts Dispersion and Uncertainty

Since the seminal work by (Bloom 2009) there is a growing literature on the measurement of the macroeconomic uncertainty and its relationship with economic activity. For instance, (Bloom 2009) used the realized volatility of SP500 (and VIX) as a proxy for macroeconomic uncertainty while (Jurado, Ludvigson & Ng 2015) (JLN) measure it as the common stochastic volatility factor of forecasting errors estimated for more than a hundred series. Another way of measuring the forecasting error volatility is to take the cross-sectional dispersion of individual forecasts for every out-of-sample period and every series for a particular forecasting horizon, see (Rossi, Sekhposyan & Soupre 2016).

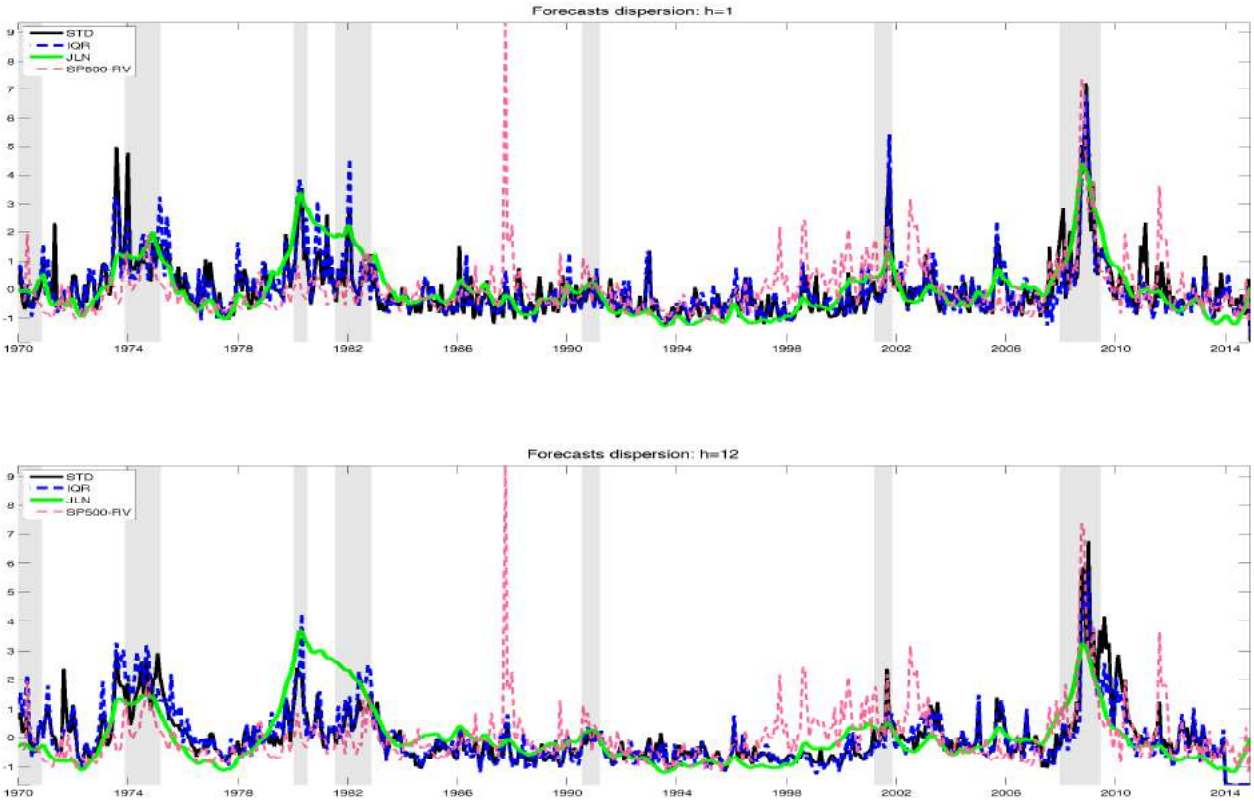
Here we consider two measures of dispersion: the standard deviation (STD) and the interquartile range (IQR). Figure 11 plots the average (across seven series) of STD and IQR against the JLN macroeconomic uncertainty measure for 1 and 12 months ahead, as well as and SP500 realized volatility. We see that the out-of-sample forecast dispersion co-moves with the macroeconomic uncertainty during the business cycles irrespective of the forecasting horizon. It increases during NBER recessions, except for the 1991 recession, and the peak dispersion is observed in the middle of the 2007-09 recession. Compared to JLN our measures present higher peaks during recessions but the two are fairly correlated.

Table 1 reports the proportion of variance of several uncertainty measures explained by our two aggregate dispersion measures and the time series-specific forecast dispersions. For instance, the aggregate STD dispersion explains 54% of variation in JLN macro uncertainty at 12-month horizon while SP500 realized volatility and VIX are explained up to 20 and 24% respectively. We also consider the economic policy uncertainty (Policy) of (Baker, Bloom & Davis 2015) with the highest R^2 of 0.20.

Finally, we verify whether the uncertainty that is measured by our out-of-sample forecast dispersion has a significant impact on the business cycle. We consider the 8 variables VAR from (Bloom 2009) and (Jurado et al. 2015) with the same recursive ordering but replacing their series of uncertainty by our aggregate STD dispersion. Figure 12 plots the impulse responses to the 100 basis points shock on forecast dispersion equation. This increase in the forecast dispersion generates a significant and persistent fall in employment and industrial production as well as in consumer prices. The federal funds rate decreases, which can be interpreted as the systematic response of the central bank. Worked hours decline in the short term. These results are in line with the findings of (Bloom 2009) and (Jurado et al. 2015).

Overall, our out-of-sample forecasts dispersion measures are good predictors of the macroeconomic and financial uncertainty measures used in the literature. An unanticipated shock to forecast dispersion can generate business cycle movements among several real activity variables.

Figure 11: Average forecasts dispersion and macroeconomic uncertainty



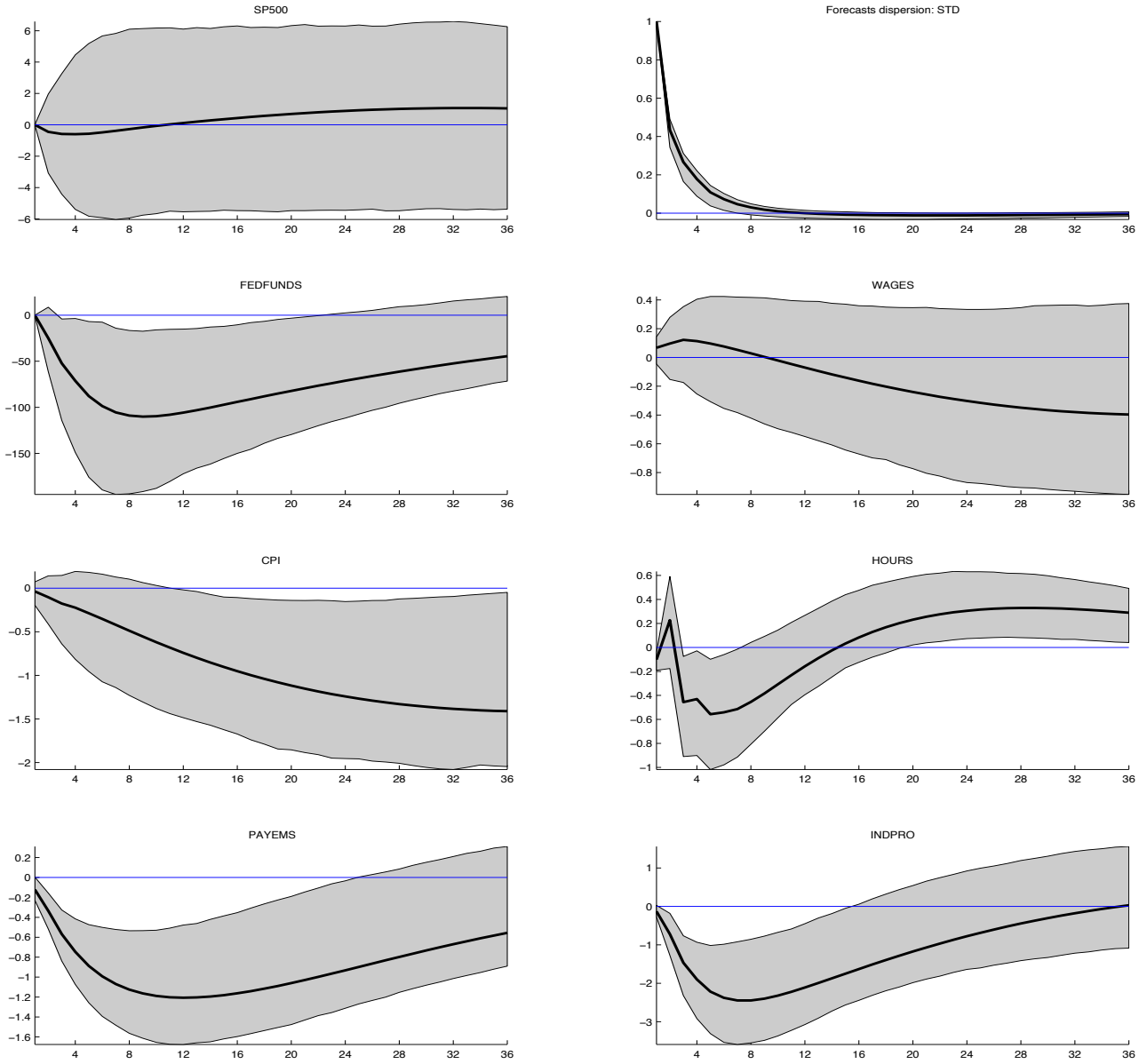
The figure shows the forecasts dispersion averaged across all 7 series for forecasting horizons of 1 and 12 months. Two dispersion measures are presented: standard error (STD) and interquartile range (IQR). JLN is the macro uncertainty from (Jurado et al. 2015) and SP500-RV is the realized volatility of SP500. All series are standardized.

Table 1: Forecasts dispersions and measures of uncertainty

		$h=1$				$h=3$				$h=12$			
		Macro	SP500	VIX	Policy	Macro	SP500	VIX	Policy	Macro	SP500	VIX	Policy
All series	STD	0,47	0,15	0,24	0,13	0,52	0,19	0,25	0,13	0,54	0,20	0,24	0,20
	IQR	0,51	0,14	0,21	0,08	0,56	0,18	0,23	0,10	0,53	0,15	0,23	0,16
INDPRO	STD	0,33	0,07	0,15	0,05	0,44	0,09	0,22	0,11	0,52	0,08	0,23	0,13
	IQR	0,27	0,05	0,13	0,04	0,33	0,07	0,19	0,09	0,32	0,04	0,15	0,13
EMP	STD	0,21	0,01	0,09	0,02	0,34	0,05	0,28	0,07	0,41	0,07	0,31	0,15
	IQR	0,19	0,02	0,06	0,01	0,24	0,04	0,17	0,03	0,25	0,03	0,20	0,08
CPI	STD	0,32	0,14	0,10	0,06	0,36	0,17	0,11	0,09	0,34	0,16	0,13	0,08
	IQR	0,21	0,11	0,04	0,02	0,28	0,15	0,07	0,03	0,34	0,15	0,11	0,05
SP500	STD	0,38	0,14	0,22	0,13	0,42	0,16	0,24	0,12	0,38	0,16	0,20	0,19
	IQR	0,42	0,12	0,19	0,08	0,47	0,15	0,23	0,09	0,38	0,13	0,20	0,15

Note: This table shows the proportion of the variance (R^2) of uncertainty measures (columns) explained by the forecasts dispersion average measures STD and IQR for horizons 1, 3 and 12 months ahead. The uncertainty measures are: Macro *uncertainty from (Jurado et al. 2015), implied volatility of SP500 index options* VIX, SP500-RV *realized volatility (measured as a standard deviation of daily returns for each month) and economic policy uncertainty from (Baker et al. 2015).*

Figure 12: Impulse responses to the shock on forecasts dispersion



This figure plots the impulse responses to the orthogonalized shock on forecast dispersion equation in the VAR-8 *model as in (Jurado et al. 2015). The lag order is set to 2 according to BIC. The gray represent 90% bootstrap confidence bands.*

9 Miscellaneous

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.

9.1 Great Recession

Figure 13 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. This may be a sign of jobless recovery that can be useful in a real-time application.[30]

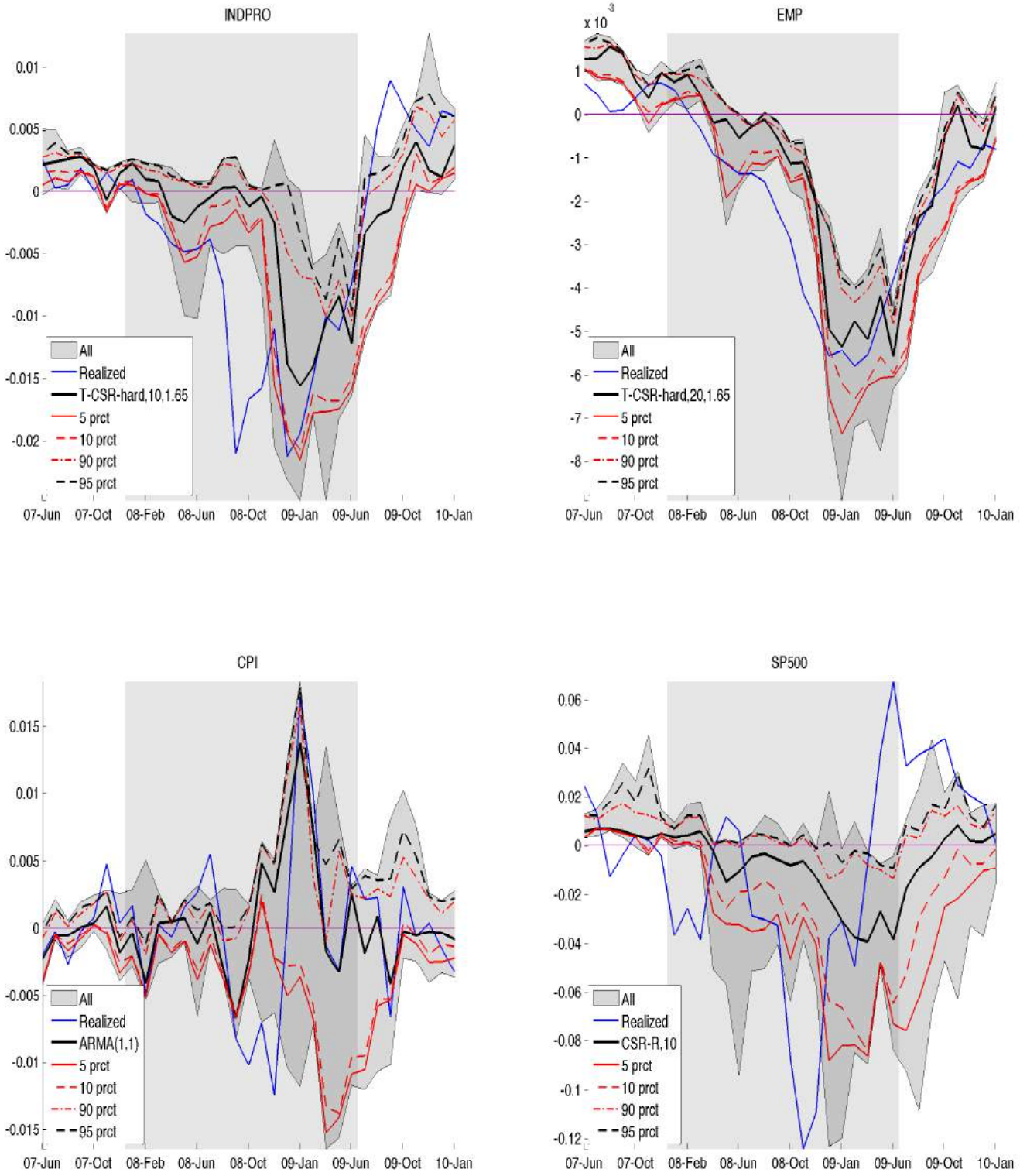
In the case of CPI inflation, ARMA forecasts track the realized values well most of the time while the low percentiles are too pessimistic from 2009 onward. The pessimistic scenarios predict downturns in stock markets, but the fast recovery at the end of the recession is not well predicted even by the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

Figure 14 plots the 3-month ahead out-of-sample 90% interval predictions of the best MSPE models as well as the distribution of all forecasts.[31] 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. This is particularly important around the business cycle turning points where the lower percentiles as well as the dispersion of forecasts are quite informative.

[30] Of course, our forecasts distribution does not come from structural models and this evidence is only suggestive.

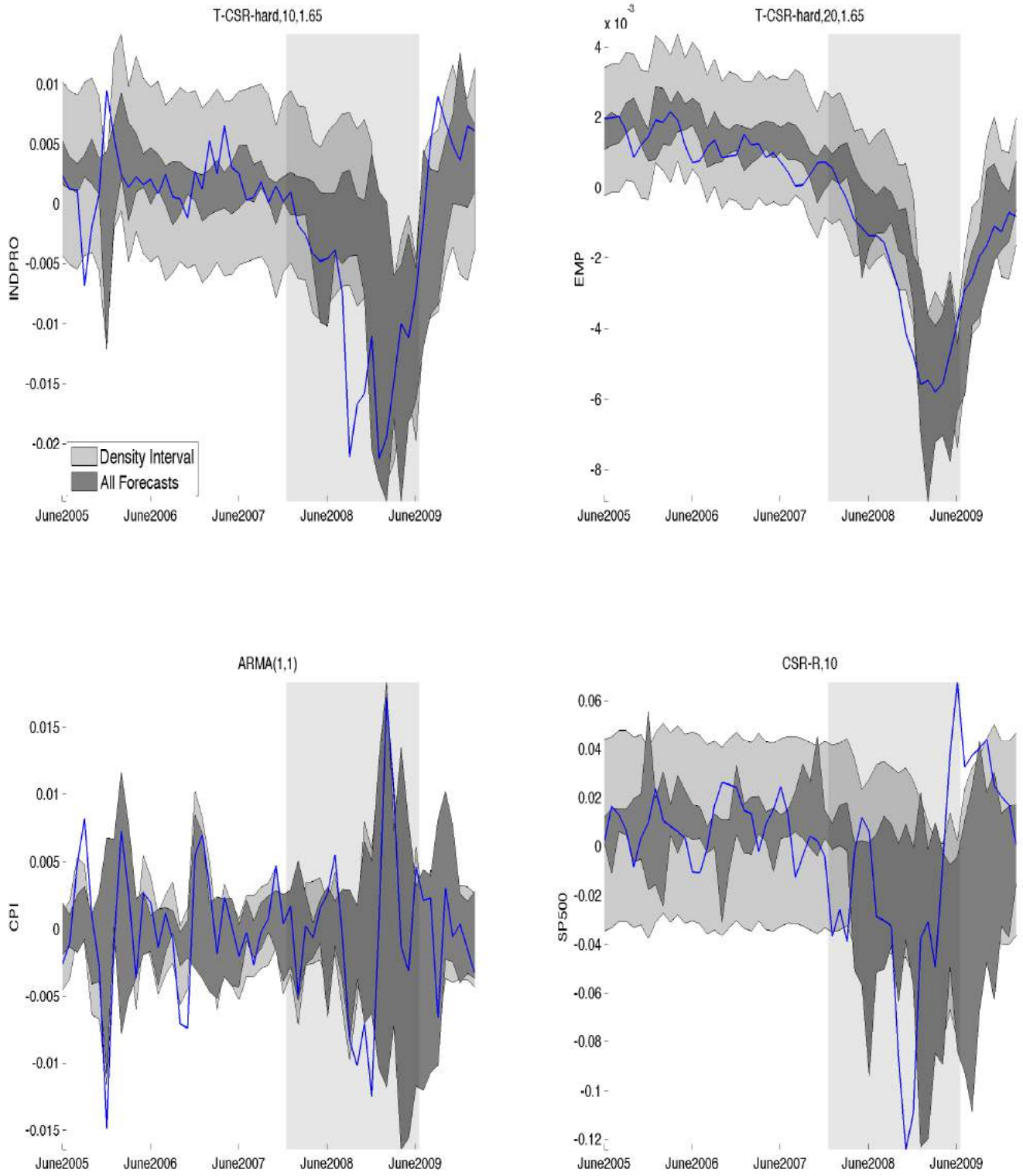
[31] Results are similar for 70% interval forecasts.

Figure 13: Forecasting 3-month ahead during Great Recession



The figure shows the 3-month ahead pseudo-out-of-sample forecasts during the Great Recession. The bleu 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.

Figure 14: Forecasting 3-month ahead during Great Recession: Interval Forecasts

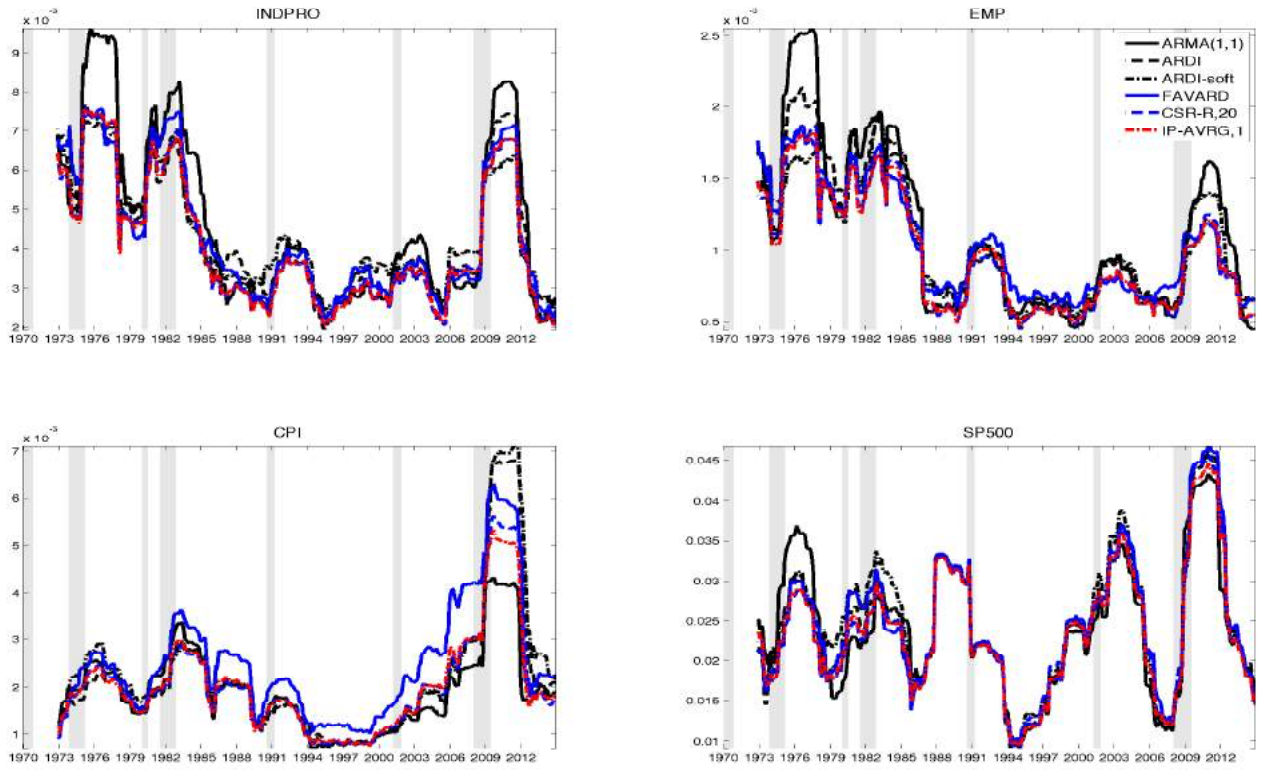


The figure shows the 3-month ahead pseudo-out-of-sample 90% interval forecasts of the best MSPE models during the Great Recession. The dark grey area represents the distribution of all 31 forecasts.

9.2 Stability of Forecast Performance

Here we verify the stability of the forecast accuracy in our pseudo-out-of-sample exercise.[32] Figures 15 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.

Figure 15: Root MSPE over time



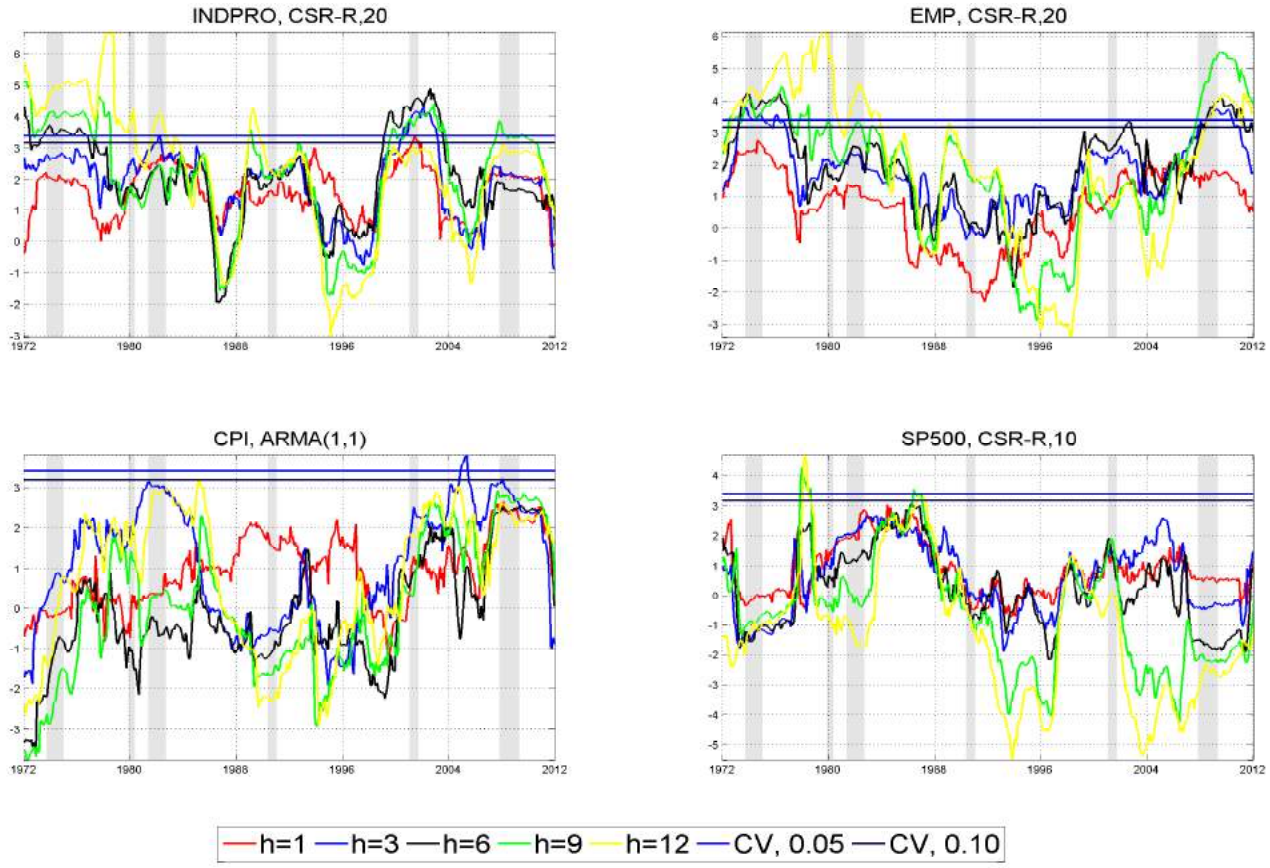
The figure shows the 3-year moving average of the root MSPE of selected models for 3-month ahead horizon.

(Giacomini & 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 16 shows the results of the Giacomini-Rossi fluctuation test for several horizons and two critical values.

[32] See (Giacomini & Rossi 2009), (Rossi & Sekhposyan 2010) and (Rossi & Sekhposyan 2011), among others, for recent examples of the time-varying forecast performance.

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 16: Giacomini-Rossi fluctuation test: Inversely proportional average



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.

9.3 Stability of Forecast Relationships

Several recent studies have suggested that factor loadings and the number of factors are likely to change over time.[33]. 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.

[33]See, among others, (Breitung & Eickmeier 2011), (D'Agostino, Gambetti & Giannone 2013), (Eickmeier, Lemke & Marcellino 2015), (Cheng, Liao & Schorfheide 2016), (Mao Takongmo & Stevanovic 2015), (Stevanovic 2016) and (Guerin, Leiva-Leon & Marcellino 2016).

In general, forecasting real activity measures require more factors (and their lags) than when predicting inflation and stock returns.[34]

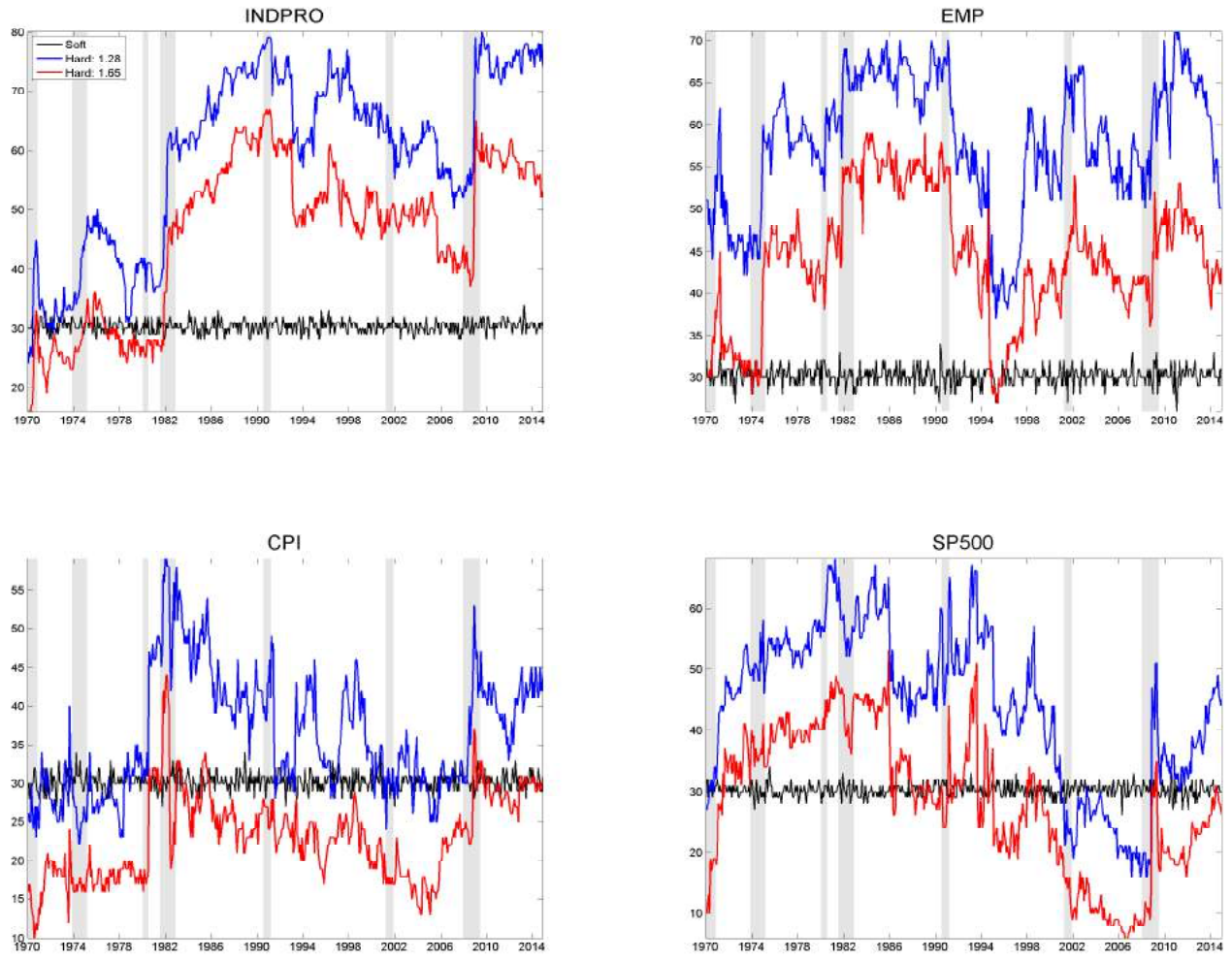
Figure 17 plots the number of series selected by soft (Lasso) and hard thresholds for all series at 3-month horizon. Recall that this is the first step in ARDIT models as well as in our targeted CSR model. The patterns of the two real activity series are quite similar. The number of candidate predictors 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 18 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 & 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.

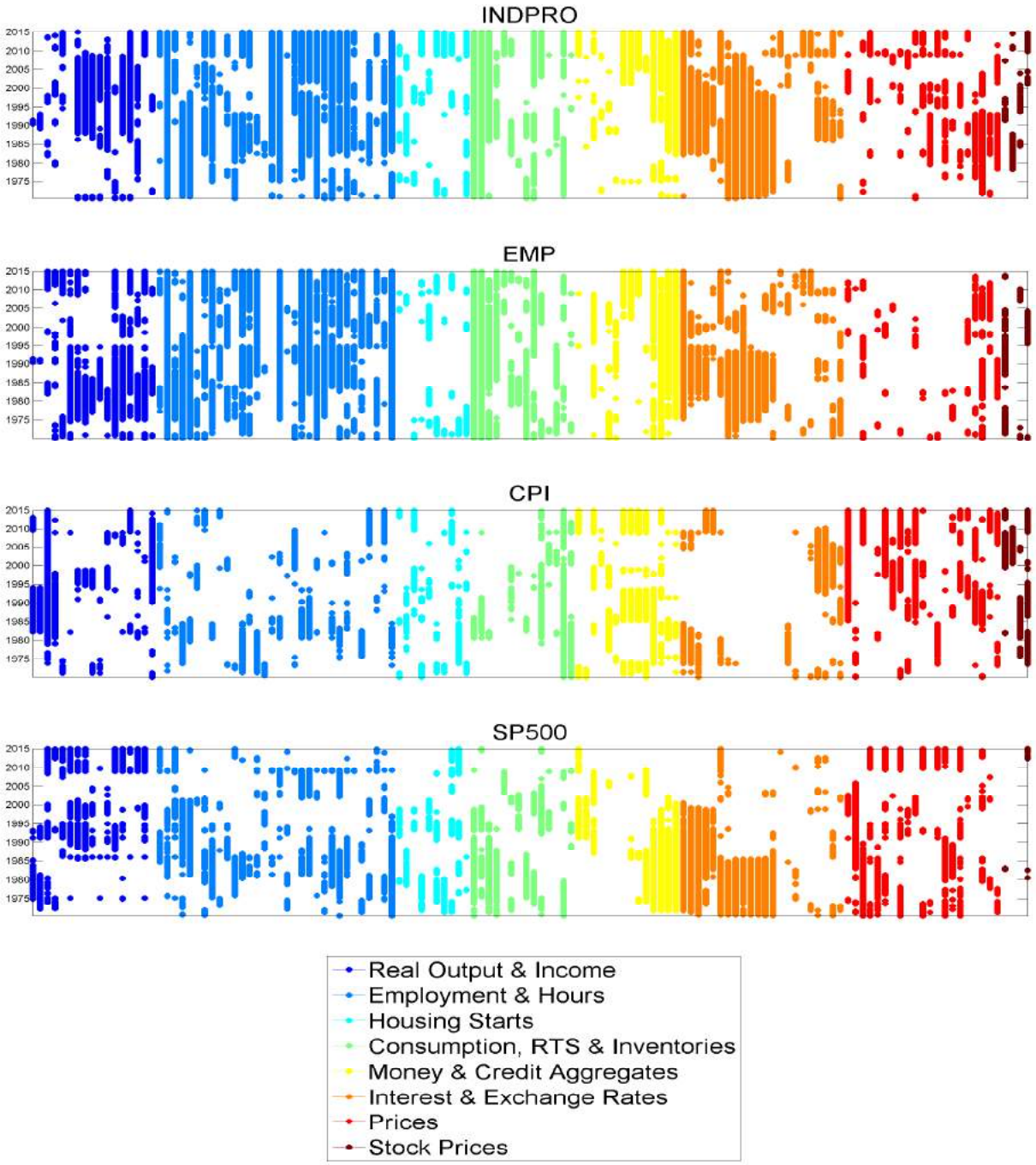
[34]For the sake of space the figures are presented in the supplementary material.

Figure 17: 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.

Figure 18: Series pre-selected by hard thresholding



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 & Ng 2015).

10 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. 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 now emerge among the best to predict real series due to their flexibility. Forecast combinations and Regularized Data-Rich models averaging techniques still perform very well relatively to the best benchmark along each performance evaluation 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. Also, we find that the dispersion of out-of-sample point forecasts is highly correlated with some macroeconomic and financial uncertainty measures used in the literature. Our study generated a huge amount of additional results that are deferred to the supplementary material.

References

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bai, J. & Ng, S. (2008), 'Forecasting economic time series using targeted predictors', *Journal of Econometrics* **146**, 304–317.

Baker, S. R., Bloom, N. & Davis, S. J. (2015), Measuring economic policy uncertainty, Technical report, NBER Working Paper No. 21633.

Banbura, M., Giannone, D. & Reichlin, L. (2010), 'Large bayesian vector autoregressions', *Journal of Applied Econometrics* **25**, 71–92.

Banerjee, A., Marcellino, M. & Masten, I. (2014), 'Forecasting with factor-augmented error correction models', *International Journal of Forecasting* **30**(3), 589–612.

Barigozzi, M., Lippi, M. & Luciani, M. (2016), Non-stationary dynamic factor models for large-datasets, Technical report, FEDS 2016-024, Board of Gov. of Federal Reserve System.

Bedock, N. & Stevanovic, D. (2016), 'An empirical study of credit shock transmission in a small open economy', *Canadian Journal of Economics* .

Bernanke, B., Boivin, J. & Eliasz, P. (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach', *Quarterly Journal of Economics* **120**, 387–422.

Bloom, N. (2009), 'The impact of uncertainty shocks', *Econometrica* **77**(3), 623–685.

Boivin, J., Giannoni, M. & Stevanović, D. (2013), Dynamic effects of credit shocks in a data-rich environment, Technical report, Federal Reserve Bank of New York Staff Reports 615.

Boivin, J. & Ng, S. (2005), 'Understanding and comparing factor-based forecasts', *International Journal of Central Banking* **1**, 117–151.

Boivin, J. & Ng, S. (2006), 'Are more data always better for factor analysis', *Journal of Econometrics* **132**, 169–194.

Breitung, J. & Eickmeier, S. (2011), 'Testing for structural breaks in dynamic factor models', *Journal of Econometrics* **163**(1), 71–84.

Carrasco, M. & Rossi, B. (2016), 'In-sample inference and forecasting in misspecified factor models', *Journal of Business and Economic Statistics* **34**(3), 313–338.

Carriero, A., Clark, T. & Marcellino, M. (2015), 'Bayesian vars: Specification choices and forecast accuracy', *Journal of Applied Econometrics* **30**, 46–73.

Cheng, X. & Hansen, B. E. (2015), 'Forecasting with factor-augmented regression: A frequentist model averaging approach', *Journal of Econometrics* **186**(2), 280–293.

Cheng, X., Liao, Z. & Schorfheide, F. (2016), 'Shrinkage estimation of high-dimensional factor models with structural instabilities', *Review of Economic Studies* **Forthcoming**.

Chevillon, G. (2007), 'Direct multi-step estimation and forecasting', *Journal of Economic Surveys* **21**(4), 746–785.

D'Agostino, A., Gambetti, L. & Giannone, D. (2013), 'Macroeconomic forecasting and structural change', *Journal of Applied Econometrics* **28**.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.

Diebold, F. X. & Pauly, P. (1987), 'Structural change and the combination of forecasts', *Journal of Forecasting* **6**, 21–40.

Dufour, J.-M. & Stevanovic, D. (2013), 'Factor-augmented VARMA models with macroeconomic applications', *Journal of Business and Economic Statistics* **31**(4), 491–506.

Eickmeier, S., Lemke, W. & Marcellino, M. (2015), 'Classical time varying factor-augmented vector auto-regressive models estimation, forecasting and structural analysis', *Journal of the Royal Statistical Society Series A* **178**(3), 493–533.

Elliott, G., Gargano, A. & Timmermann, A. (2013), 'Complete subset regressions', *Journal of Econometrics* **177**(2), 357–373.

Faust, J. & Wright, J. (2013), Forecasting inflation, *in* G. Elliott & A. Timmermann, eds, 'Handbook of Economic Forecasting', Vol. 2A, Elsevier.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2005), 'The generalized dynamic factor model: one-sided estimation and forecasting', *Journal of the American Statistical Association* **100**, 830–839.

Foroni, C., Marcellino, M. & Stevanovic, D. (2017), Mixed frequency models with MA components, Technical report, Department of Economics, UQAM.

Galbraith, J. (2003), 'Content horizons for univariate time series forecasts', *International Journal of Forecasting* **19**(1), 43–55.

Giacomini, R. & Rossi, B. (2009), 'Detecting and predicting forecast breakdowns', *Review of Economic Studies* **76**(2).

Giacomini, R. & Rossi, B. (2010), 'Forecast comparisons in unstable environments', *Journal of Applied Econometrics* **25**(4), 595 – 620.

Giannone, D., Lenza, M. & Primiceri, G. (2015), 'Prior selection for vector autoregressions', *Review of Economics and Statistics* **97**(2), 436–451.

Giannone, D., Lenza, M. & Primiceri, G. (2017), Macroeconomic prediction with big data: the illusion of sparsity, Technical report, Federal Reserve Bank of New York.

Groen, J. J. & Kapetanios, G. (2016), 'Revisiting useful approaches to data-rich macroeconomic forecasting', *Computational Statistics and Data Analysis* **100**, 221–239.

Guerin, P., Leiva-Leon, D. & Marcellino, M. (2016), Markov-switching three-pass regression filter, Technical report, Department of Economics, Bocconi.

Hallin, M. & Liska, R. (2007), 'Determining the number of factors in the general dynamic factor model', *Journal of the American Statistical Association* **102**, 603–617.

Inoue, A., Kilian, L. & Rossi, B. (2016), 'Optimal window selection in the presence of possible instabilities', *Journal of Econometrics* **Forthcoming**.

Jurado, K., Ludvigson, S. & Ng, S. (2015), 'Measuring uncertainty', *American Economic Review* **105**(3), 1177–1216.

Kelly, B. & Pruitt, S. (2015), 'The three-pass regression filter: A new approach to forecasting using many predictors', *Journal of Econometrics* **186**(2), 294–316.

Kim, H. H. & Swanson, N. R. (2014), 'Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence', *Journal of Econometrics* **178**(2), 352–367.

Koop, G. (2013), 'Forecasting with medium and large bayesian vars', *Journal of Applied Econometrics* **28**, 177–203.

Mao Takongmo, C. & Stevanovic, D. (2015), 'Selection of the number of factors in presence of structural instability: a monte carlo study', *Actualit conomique* **91**, 177–233.

Marcellino, M., Stock, J. H. & Watson, M. W. (2006), 'A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series', *Journal of Econometrics* **135**, 499–526.

McCracken, M. W. & Ng, S. (2015), 'Fred-md: A monthly database for macroeconomic research', *Journal of Business and Economic Statistics* .

Ng, S. & Perron, P. (1996), 'Useful modifications to some unit root tests in with dependent errors and their local asymptotic properties', *Review of Economic Studies* **63**, 435–463.

Ng, S. & Perron, P. (2001), 'Lag length selection and the construction of unit root tests with good size and power', *Econometrica* **69**(6), 1519–1554.

Patton, A. & Timmermann, A. (2012), 'Forecast rationality tests based on multi-horizon bounds', *Journal of Business and Economic Statistics* **30**(1), 1–17.

Pesaran, H. & Timmermann, A. (1992), 'A simple nonparametric test of predictive performance', *Journal of Business and Economic Statistics* **10**(4), 461–465.

Rapach, D. E. Strauss, J. K. & Zhou, G. (2010), 'Out-of-sample equity premium prediction: Combination forecasts and links to the real economy', *Review of Financial Studies* **23**(2), 821–862.

Rossi, B. & Sekhposyan, T. (2010), 'Have models? forecasting performance changed over time, and when?', *International Journal of Forecasting* **26**(4).

Rossi, B. & Sekhposyan, T. (2011), 'Understanding models? forecasting performance', *Journal of Econometrics* **164**(1), 158–172.

Rossi, B., Sekhposyan, T. & Soupre, M. (2016), Understanding the sources of macroeconomic uncertainty, Technical report, Department of Economics, Universitat Pompeu Fabra.

Ruge-Murcia, F. & Onatski, A. (2013), 'Factor analysis of a large dsge model', *Journal of Applied Econometrics* **28**(6), 903–928.

Satchell, S. & Timmermann, A. (1995), 'An assessment of the economic value of non-linear foreign exchange rate forecasts', *Journal of Forecasting* **14**(6), 477–497.

Stevanovic, D. (2015), Factor-augmented autoregressive distributed lag model with macroeconomic applications, Technical report, Department of Economics, UQAM.

Stevanovic, D. (2016), 'Common time variation of parameters in reduced-form macroeconomic models', *Studies in Nonlinear Dynamics & Econometrics* **20**(2), 159–183.

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

Stock, J. H. & Watson, M. W. (2002b), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business and Economic Statistics* **20**(2), 147–162.

Stock, J. H. & Watson, M. W. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**, 405–430.

Stock, J. H. & Watson, M. W. (2005), Implications of dynamic factor models for var analysis, Technical report, NBER WP 11467.

Stock, J. H. & Watson, M. W. (2007), 'Why has U.S. inflation become harder to forecast?', *Journal of Money, Credit and Banking* **39**(1).