Learning from Forecast Errors: A New Approach to Forecast Combination

Tae-Hwy Lee* and Ekaterina Seregina†

September 14, 2020

Abstract

This paper studies forecast combination (as an expert system) using the precision matrix estimation of forecast errors when the latter admit the approximate factor model. This approach incorporates the facts that experts often use common sets of information and hence they tend to make common mistakes. This premise is evidenced in many empirical results. For example, the European Central Bank's Survey of Professional Forecasters on Euro-area real GDP growth demonstrates that the professional forecasters tend to *jointly* understate or overstate GDP growth. Motivated by this stylized fact, we develop a novel framework which exploits the factor structure of forecast errors and the sparsity in the precision matrix of the idiosyncratic components of the forecast errors. The proposed algorithm is called *Factor Graphical Model* (FGM). Our approach overcomes the challenge of obtaining the forecasts that contain unique information, which was shown to be necessary to achieve a "winning" forecast combination. In simulation, we demonstrate the merits of the FGM in comparison with the equal-weighted forecasts and the standard graphical methods in the literature. An empirical application to forecasting macroeconomic time series in big data environment highlights the advantage of the FGM approach in comparison with the existing methods of forecast combination.

Keywords: High-dimensionality; Graphical Lasso; Approximate Factor Model; Nodewise Regression; Precision Matrix

JEL Classifications: C13, C38, C55

*Department of Economics, University of California, Riverside. Email: tae.lee@ucr.edu.

†Department of Economics, University of California, Riverside. Email: ekaterina.seregina@email.ucr.edu.

1 Introduction

A search for the best forecast combination has been an important on-going research question in economics. Clemen (1989) pointed out that combining forecasts is "practical, economical and useful. Many empirical tests have demonstrated the value of composite forecasting. We no longer need to justify that methodology". However, as demonstrated by Diebold and Shin (2019), there are still some unresolved issues. Despite the findings based on the theoretical grounds, equal-weighted forecasts have proved surprisingly difficult to beat. It can be shown (see Timmermann (2006)) that equal weights are optimal in situations with an arbitrary number of forecasts when the individual forecast errors have the same variance and identical pairwise correlations. Many methodologies that seek for the best forecast combination use equal weights as a benchmark: for instance, Diebold and Shin (2019) develop "partially egalitarian Lasso" that discards some forecasts and then selects and shrinks the remaining forecasts toward equal weights.

In this paper we are interested in finding the combination of forecasts which yields the best out-of-sample performance in terms of the mean-squared forecast error (MSFE). We claim that the success of equal weights is partly due to the fact that the forecasters use the same set of public information to make forecasts, hence, they tend to make common mistakes. This statement is supported by the empirical results based on the European Central Bank's Survey of Professional forecasters of Euro-area real GDP growth. We demonstrate that the forecasters tend to *jointly* understate or overstate GDP growth. Therefore, we propose that the forecast errors include common and idiosyncratic components, which allows us to model the tendency of the forecast errors to move together due to the common component.

Several recent papers support this methodology: Atiya (2020) provides a good graphical illustration showing that "forecast combination should be a winning strategy if the constituent forecasts are either diverse or comparable in performance". Thomson et al. (2019) find that the benefit of combining forecasts from many models/experts depends upon the extent to which these individual forecasts contain unique information. Our paper provides a simple framework to separate unique/individual errors from the common ones to improve the accuracy of the combined forecast.

In high dimensions, when the number of forecasts is large relative to the sample size, the sample covariance matrix of the forecast errors underlying the standard forecast combination (see Bates and

Granger (1969)) is subject to the estimation uncertainty. In the literature on portfolio allocation and forecast combination, there are three popular methods to construct a good covariance matrix estimator. The first one uses shrinkage of the sample covariance matrix (see Ledoit and Wolf (2003, 2004a,b, 2012, 2015) among others). The second one imposes some structure on the data, such as using factor models to decompose covariance matrix into a low-rank and sparse components (see Fan et al. (2011, 2019, 2017)), and the third one uses thresholding of the sample covariance (see Bickel and Levina (2008); Cai and Liu (2011)). Rothman et al. (2008) emphasized that shrinkage estimators of the form proposed by Ledoit and Wolf (2003, 2004b) do not affect the eigenvectors of the covariance, only the eigenvalues. However, Johnstone and Lu (2009) showed that the sample eigenvectors are also not consistent in high-dimensions. When constructing a covariance matrix estimator using the factor model for high-dimensional problems, after we estimate the low-rank component, the idiosyncratic part is still large. Therefore, in addition to a factor model, we need some sparsity condition for estimating the residual precision matrix. The thresholding estimators result in sparse covariance but do not take into account the structure in the data. Fan et al. (2008) showed that the precision matrix takes advantage of the factor structure and, hence, can be better estimated in the factor approach. At the same time, forecast combination problem requires an estimator of the precision matrix. However, regularizing a covariance matrix does not guarantee a well-behaved estimator of the precision.

Our paper develops a new precision matrix estimator for the forecast errors under the approximate factor model with unobserved factors that addresses the aforementioned limitations. We call our algorithm the *Factor Graphical Model*. We use a factor model to estimate a sparse idiosyncratic component, and then apply the Graphical model (Graphical Lasso (Friedman et al. (2008)) or nodewise regression (Meinshausen and Bühlmann (2006))) for the estimation of the precision matrix of the idiosyncratic terms.

There are a few papers that used graphical models to estimate the covariance matrix of the idiosyncratic component when the factors are known and the loadings are assumed to be constant. Brownlees et al. (2018) estimate a sparse covariance matrix for high-frequency data and construct the realized network for financial data. Barigozzi et al. (2018) develop a power-law partial correlation network based on the Gaussian graphical models. They show that when the dimension of the system is large, the largest eigenvalues of the precision converge to a positive affine transfor-

mation. Koike (2020) uses the Weighted Graphical Lasso to estimate a sparse covariance matrix of the idiosyncratic component for a factor model with observable factors for high-frequency data. The paper derives consistency and the asymptotic mixed normality for the estimator based on the realized covariance matrix.

Our paper makes several important contributions: first, we develop a novel framework that models the tendency of the forecast errors to move together due to the common component. This framework is supported by the stylized fact that the forecasters tend to jointly understate or overstate the predicted series of interest. Second, we develop a novel high-dimensional precision matrix estimator which combines the benefits of the factor structure and sparsity of the precision matrix of the idiosyncratic component for the forecast combination under the approximate factor model. Third, this is the first paper that provides a simple framework to overcome the challenge of obtaining the forecasts that contain unique information, which was shown to be necessary to achieve a "winning" forecast combination. The empirical application highlights the advantage of this methodology in comparison with the existing methods of forecast combination.

The paper is structured as follows: Section 2 reviews Graphical Lasso and nodewise regression techniques. Section 3 studies the approximate factor models for the forecast combination. Section 4 introduces the Factor Graphical Model and discusses the tuning of the proposed model. Section 5 provides simulations. Section 6 studies an empirical application for macroeconomic time-series. And Section 7 concludes.

Notation. For the convenience of the reader, we summarize the notation to be used throughout the paper. Given a vector $\mathbf{u} \in \mathbb{R}^d$ and a parameter $a \in [1, \infty)$, let $\|\mathbf{u}\|_a$ denote l_a-norm. Given a matrix $\mathbf{U} \in \mathbb{R}^{p \times p}$ and parameters $a, b \in [1, \infty)$, let $|||\mathbf{U}|||_{a,b}$ denote the induced matrix-operator norm $\max_{\|\mathbf{y}\|_a=1} \|\mathbf{U}\mathbf{y}\|_b$. The special cases are $|||\mathbf{U}|||_1 \coloneqq \max_{1 \le j \le p} \sum_{i=1}^{p} |\mathbf{U}_{i,j}|$ for the l_1/l_1-operator norm; the operator norm (l_2-matrix norm) $|||\mathbf{U}|||_2^2 \coloneqq \Lambda_{max}(\mathbf{U}\mathbf{U}')$ is equal to the maximal singular value of $\mathbf{U}$. Finally, $\|\mathbf{U}\|_\infty$ denotes the element-wise maximum $\max_{i,j} |\mathbf{U}_{i,j}|$.

2 Gaussian Graphical Models for Forecast Errors

This section briefly reviews a class of models, called Gaussian graphical models, that search for the estimator of the precision matrix (see Bishop (2006); Hastie et al. (2001) for more detailed description). In graphical models, each vertex represents a random variable, and the graph visualizes

the joint distribution of the entire set of random variables.

Sparse graphs have a relatively small number of edges. Among the main challenges in working with the graphical models are choosing the structure of the graph (*model selection*) and estimation of the edge parameters from the data.

Suppose we have p competing forecasts of the univariate series y_t, $t = 1, \ldots, T$. Let $\mathbf{e}_t = (e_{1t}, \ldots, e_{pt})' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ be a $p \times 1$ vector of forecast errors. Assume they follow a Gaussian distribution. The precision matrix $\mathbf{\Sigma}^{-1} \coloneqq \mathbf{\Theta}$ contains information about partial covariances between the variables. For instance, if $\mathbf{\Theta}_{ij}$, which is the ij-th element of the precision matrix, is zero, then the variables i and j are conditionally independent, given the other variables.

Given a sample $\{\mathbf{e}_t\}_{t=1}^T$, let $\mathbf{S} = (1/T)\sum_{t=1}^T (\mathbf{e}_t)(\mathbf{e}_t)'$ denote the sample covariance matrix and $\widehat{\mathbf{D}}^2 \coloneqq \text{diag}(\mathbf{S})$. We can write down the Gaussian log-likelihood (up to constants) $l(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \text{trace}(\mathbf{S\Theta})$. The maximum likelihood estimate (MLE) of $\mathbf{\Theta}$ is $\widehat{\mathbf{\Theta}} = \mathbf{S}^{-1}$.

In the high-dimensional settings it is necessary to regularize the precision matrix, which means that some edges will be zero. In the following subsections we discuss two most widely used techniques to estimate sparse high-dimensional precision matrices.

2.1 Graphical Lasso

The first approach to induce sparsity in the estimation of precision matrix is to add penalty to the maximum likelihood and use the connection between the precision matrix and regression coefficients to maximize the following *weighted penalized log-likelihood* (Janková and van de Geer (2018)):

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}=\mathbf{\Theta}'} \text{trace}(\mathbf{S\Theta}) - \log\det(\mathbf{\Theta}) + \lambda \sum_{i \neq j} \widehat{\mathbf{D}}_{ii}\widehat{\mathbf{D}}_{jj}|\mathbf{\Theta}_{ij}|, \tag{2.1}$$

over positive definite symmetric matrices, where $\lambda \geq 0$ is a penalty parameter. When $\lambda = 0$, the MLEs for $\mathbf{\Sigma}$ and $\mathbf{\Theta}$ in (2.1) are the sample covariance matrix $\mathbf{S}$ and its inverse $\mathbf{S}^{-1}$ respectively. When $\lambda > 0$, the solution yields *penalized MLEs* of the covariance and precision matrices, denoted as $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Sigma}}^{-1}$.

The penalized likelihood formulation was proposed in Yuan and Lin (2007). They solve the optimization problem in (2.1) using the interior-point method for the max log-determinant problem (see Vandenberghe et al. (1998) for more details on the method). This procedure guarantees

the positive-definiteness of the penalized MLE of $\mathbf{\Theta}$. However, the method is computationally demanding and is limited to $p \leq 10$. Banerjee et al. (2008) develop a different framework using block-coordinate descent. They show that the optimization problem in (2.1) is convex and consider estimation of $\mathbf{\Sigma}$ rather than $\mathbf{\Theta}$. Banerjee et al. (2008) show that one can easily recover $\mathbf{\Theta}$ using their procedure.

One of the most popular and fast algorithms to solve the optimization problem in (2.1) is called the Graphical Lasso (GLASSO), which was introduced by Friedman et al. (2008). The Graphical Lasso procedure is summarized in Algorithm 1: it combines the neighborhood method by Meinshausen and Bühlmann (2006) and block-coordinate descent by Banerjee et al. (2008).

Algorithm 1 Graphical Lasso (Friedman et al. (2008))

1: Initialize $\mathbf{W} = \mathbf{S} + \lambda \mathbf{I}$. The diagonal of $\mathbf{W}$ remains the same in what follows.
2: Repeat for $j = 1, \ldots, p, 1, \ldots, p, \ldots$ until convergence:

- Partition $\mathbf{W}$ into part 1: all but the j-th row and column, and part 2: the j-th row and column,
- Solve the score equations using the cyclical coordinate descent:

$$\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} + \lambda \cdot \text{Sign}(\boldsymbol{\beta}) = \mathbf{0}.$$

  This gives a $(p-1) \times 1$ vector solution $\widehat{\boldsymbol{\beta}}$.
- Update $\widehat{\mathbf{w}}_{12} = \mathbf{W}_{11}\widehat{\boldsymbol{\beta}}$.

3: In the final cycle (for $i = 1, \ldots, p$) solve for

$$\frac{1}{\widehat{\theta}_{22}} = w_{22} - \widehat{\boldsymbol{\beta}}' \widehat{\mathbf{w}}_{12}, \quad \widehat{\boldsymbol{\theta}}_{12} = -\widehat{\theta}_{22}\widehat{\boldsymbol{\beta}}.$$

2.2 Nodewise Regression

An alternative approach to induce sparsity in the estimation of precision matrix in equation (2.1) is to solve for $\widehat{\mathbf{\Theta}}$ one column at a time via linear regressions, replacing population moments by their sample counterparts $\mathbf{S}$. When we repeat this procedure for each variable $j = 1, \ldots, p$, we will estimate the elements of $\widehat{\mathbf{\Theta}}$ column by column using $\{\mathbf{e}_t\}_{t=1}^T$ via p linear regressions. Meinshausen and Bühlmann (2006) use this approach (which we will refer to as MB) to incorporate sparsity into the estimation of the precision matrix. They fit p separate Lasso regressions using each variable (node) as the response and the others as predictors to estimate $\widehat{\mathbf{\Theta}}$. This method is known as the

"nodewise" regression and it is reviewed below based on van de Geer et al. (2014) and Callot et al. (2019).

Let $\mathbf{e}_j$ be a $T \times 1$ vector of observations for the j-th regressor, the remaining covariates are collected in a $T \times p$ matrix $\mathbf{E}_{-j}$. For each $j = 1, \ldots, p$ we run the following Lasso regressions:

$$\widehat{\boldsymbol{\gamma}}_j = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \left(\|\mathbf{e}_j - \mathbf{E}_{-j}\|_2^2 / T + 2\lambda_j \|\boldsymbol{\gamma}\|_1 \right), \tag{2.2}$$

where $\widehat{\boldsymbol{\gamma}}_j = \{\widehat{\gamma}_{j,k}; j = 1, \ldots, p, k \neq j\}$ is a $(p-1) \times 1$ vector of the estimated regression coefficients that will be used to construct the estimate of the precision matrix, $\widehat{\boldsymbol{\Theta}}$. Define

$$\widehat{\mathbf{C}} = \begin{pmatrix} 1 & -\widehat{\gamma}_{1,2} & \cdots & -\widehat{\gamma}_{1,p} \\ -\widehat{\gamma}_{2,1} & 1 & \cdots & -\widehat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\gamma}_{p,1} & -\widehat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}. \tag{2.3}$$

For $j = 1, \ldots, p$, define

$$\hat{\tau}_j^2 = \|\mathbf{e}_j - \mathbf{E}_{-j}\widehat{\boldsymbol{\gamma}}_j\|_2^2 / T + \lambda_j \|\widehat{\boldsymbol{\gamma}}_j\|_1 \tag{2.4}$$

and write

$$\widehat{\mathbf{T}}^2 = \text{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_p^2). \tag{2.5}$$

The approximate inverse is defined as

$$\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}. \tag{2.6}$$

One of the caveats to keep in mind when using the MB method is that the estimator in (2.6) is not self-adjoint. Callot et al. (2019) show (see their Lemma A.1) that $\widehat{\boldsymbol{\Theta}}$ in (2.6) is positive definite with high probability, however, it could still occur that $\widehat{\boldsymbol{\Theta}}$ is not positive definite in finite samples. A possible solution would be to use the matrix symmetrization procedure as in Fan et al. (2018) and then use eigenvalue cleaning as in Callot et al. (2017) and Hautsch et al. (2012). The procedure to estimate the precision matrix using nodewise regression is summarized in Algorithm 2.

Algorithm 2 Nodewise regression by Meinshausen and Bühlmann (2006) (MB)

1: Repeat for $j = 1, \ldots, p$:

- Estimate $\widehat{\boldsymbol{\gamma}}_j$ using (2.2) for a given λ_j.
- Select λ_j using a suitable information criterion (see section 4.1 for the possible options).

2: Calculate $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{T}}^2$.
3: Return $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{T}}^{-2}\widehat{\mathbf{C}}$.

3 Approximate Factor Models for Forecast Errors

The approximate factor models for the forecasts were first considered by Chan et al. (1999). They modeled a panel of ex-ante forecasts of a single time-series as a dynamic factor model and found out that the combined forecasts improved on individual ones when all forecasts have the same information set (up to difference in lags). This result emphasizes the benefit of forecast combination even when the individual forecasts are not based on different information and, therefore, do not broaden the information set used by any one forecaster.

In this paper, we are interested in finding the combination of forecasts which yields the best out-of-sample performance in terms of the mean-squared forecast error to be introduced later. We claim that the forecasters use the same set of public information to make forecasts, hence, they tend to make common mistakes. Figure 1 illustrates this statement: it shows quarterly forecasts of Euro-area real GDP growth produced by the European Central Bank's Survey of Professional Forecasters from 1999Q3 to 2019Q3. As described in Diebold and Shin (2019), forecasts are solicited for one year ahead of the latest available outcome: e.g., the 2007Q1 survey asked the respondents to forecast the GDP growth over 2006Q3-2007Q3. As evidenced from Figure 1, forecasters tend to jointly understate or overstate GDP growth, meaning that their forecast errors include common and idiosyncratic parts. Therefore, we can model the tendency of the forecast errors to move together via factor decomposition.

Recall that we have p competing forecasts of the univariate series y_t, $t = 1, \dots, T$ and $\mathbf{e}_t = (e_{1t}, \dots, e_{pt})' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is a $p \times 1$ vector of forecast errors. Assume that the generating process for the forecast errors follows a q-factor model:

$$\underbrace{\mathbf{e}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{q\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, T \tag{3.1}$$

where $\mathbf{f}_t = (f_{1t}, \dots, f_{qt})'$ are the factors of the forecast errors for p models, $\mathbf{B}$ is a $p \times q$ matrix of factor loadings, and $\boldsymbol{\varepsilon}_t$ is the idiosyncratic component that cannot be explained by the common factors. Unobservable factors and loadings are usually estimated by the principal component analysis (PCA), studied in Bai (2003); Bai and Ng (2002); Connor and Korajczyk (1988); Stock and Watson (2002). Strict factor structure assumes that the idiosyncratic disturbances, $\boldsymbol{\varepsilon}_t$, are uncorrelated with each other, whereas approximate factor structure allows correlation of the idiosyncratic

disturbances (Chamberlain and Rothschild (1983)).

We use the following notations: $\mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}_\varepsilon$, $\mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}_\varepsilon$, $\mathbb{E}[\mathbf{e}_t\mathbf{e}_t'] = \boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Sigma}_f\mathbf{B}' + \boldsymbol{\Sigma}_\varepsilon$, and $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathbf{f}_t] = 0$. The objective function to recover factors and loadings from (3.1) is:

$$\min_{\mathbf{f}_1,\ldots,\mathbf{f}_T,\mathbf{B}} \frac{1}{T}\sum_{t=1}^{T}(\mathbf{e}_t - \mathbf{B}\mathbf{f}_t)'(\mathbf{e}_t - \mathbf{B}\mathbf{f}_t) \tag{3.2}$$

$$\text{s.t. } \mathbf{B}'\mathbf{B} = \mathbf{I}_q, \tag{3.3}$$

where (3.3) is the assumption necessary for the unique identification of factors. Fixing the value of $\mathbf{B}$, we can project forecast errors $\mathbf{e}_t$ into the space spanned by $\mathbf{B}$: $\mathbf{f}_t = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{e}_t = \mathbf{B}'\mathbf{e}_t$. When combined with (3.2), this yields a concentrated objective function for $\mathbf{B}$:

$$\max_{\mathbf{B}} \ \text{tr}\Big[\mathbf{B}'\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t'\Big)\mathbf{B}\Big]. \tag{3.4}$$

It is well-known (see Stock and Watson (2002) among others) that $\widehat{\mathbf{B}}$ estimated from the first q eigenvectors of $\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t'$ is the solution to (3.4). Once we obtain $\widehat{\mathbf{f}}_t, \widehat{\mathbf{B}}$, we can get an estimate of covariance matrix of forecast errors $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{B}}\widehat{\boldsymbol{\Sigma}}_f\widehat{\mathbf{B}}' + \widehat{\boldsymbol{\Sigma}}_\varepsilon$.

Having obtained all necessary estimates, we move to the forecast combination exercise. Suppose we have p competing forecasts, $\widehat{\mathbf{y}}_t = (\hat{y}_{1,t},\ldots,\hat{y}_{p,t})'$, of the variable y_t, $t = 1,\ldots,T$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix of the forecast errors. The forecast combination is defined as follows:

$$\widehat{y}_t^c = \mathbf{w}'\widehat{\mathbf{y}}_t \tag{3.5}$$

where $\mathbf{w}$ is a $p \times 1$ vector of weights. Define a measure of risk $R(\mathbf{w}, \boldsymbol{\Sigma}) = \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$. As shown in Bates and Granger (1969), the *optimal* forecast combination minimizes the variance of the combined forecast error:

$$\min_{\mathbf{w}} R(\mathbf{w}, \boldsymbol{\Sigma}), \ \text{s.t. } \mathbf{w}'\boldsymbol{\iota}_p = 1, \tag{3.6}$$

where $\boldsymbol{\iota}_p$ is a $p \times 1$ vector of ones. The solution to (3.6) yields a $p \times 1$ vector of the optimal forecast combination weights:

$$\mathbf{w} = \frac{\boldsymbol{\Theta}\boldsymbol{\iota}_p}{\boldsymbol{\iota}_p'\boldsymbol{\Theta}\boldsymbol{\iota}_p}. \tag{3.7}$$

If the true precision matrix is known, the equation (3.7) guarantees to yield the optimal forecast combination. In reality, one has to estimate $\boldsymbol{\Theta}$. Hence, the out-of-sample performance of the combined forecast is affected by the estimation error. As pointed out by Smith and Wallis (2009)

and Claeskens et al. (2016), when the estimation uncertainty of the weights is taken into account, there is no guarantee that the "optimal" forecast combination will be better than the equal weights or even improve the individual forecasts. Concretely, for any estimator of covariance matrix and combination weights, we have:

$$\left|R(\widehat{\mathbf{w}},\widehat{\mathbf{\Sigma}})-R(\mathbf{w},\widehat{\mathbf{\Sigma}})\right|\leq\|\widehat{\mathbf{w}}-\mathbf{w}\|_1\left\|\widehat{\mathbf{\Sigma}}\mathbf{w}\right\|_\infty. \tag{3.8}$$

The equation (3.8) implies that given an estimator of the covariance matrix, the risk of the combined forecast is bounded by the estimation error in the optimal forecast combination weights. In order to control the latter, define $a=\boldsymbol{\iota}_p'\mathbf{\Theta}\boldsymbol{\iota}_p/p$, and $\widehat{a}=\boldsymbol{\iota}_p'\widehat{\mathbf{\Theta}}\boldsymbol{\iota}_p/p$. We can easily obtain the following bound on the optimal combination weights:

$$\|\widehat{\mathbf{w}}-\mathbf{w}\|_1\leq\frac{a\frac{\left\|(\widehat{\mathbf{\Theta}}-\mathbf{\Theta})\boldsymbol{\iota}_p\right\|_1}{p}+|a-\widehat{a}|\frac{\|\mathbf{\Theta}\boldsymbol{\iota}_p\|_1}{p}}{|\widehat{a}|a}, \tag{3.9}$$

where the inequality was shown in Callot et al. (2019). Therefore, in order to control the estimation uncertainty in the combination weights, one needs to obtain a consistent estimator of the precision matrix $\mathbf{\Theta}$.

4 Factor Graphical Models for Forecast Errors

We can use the optimization problem in (2.1) or (2.2) to directly estimate the precision matrix $\mathbf{\Theta}$ and apply it to find the forecast combination weights in (3.7). As pointed out by Dai et al. (2019), for high-dimensional problems after we estimate the low-rank component of the covariance matrix, the idiosyncratic part is still large. Therefore, in addition to a factor model, we need some sparsity condition for estimating the residual covariance matrix, $\mathbf{\Sigma}_\varepsilon$. Rothman et al. (2008) emphasized that shrinkage estimators of the form proposed by Ledoit and Wolf (2003, 2004b) do not affect the eigenvectors of the covariance, only the eigenvalues. However, Johnstone and Lu (2009) showed that the sample eigenvectors are also not consistent in high-dimensions. Fan et al. (2011) first construct a sample covariance matrix of residuals based on the estimated factor model, and then apply adaptive thresholding to estimate the idiosyncratic component of the covariance matrix.

Since our interest is in constructing weights for the forecast combination, our goal is to estimate a precision matrix of the forecast errors. However, as pointed out by Koike (2020), when common

factors are present across the forecast errors, the precision matrix cannot be sparse because all pairs of the forecast errors are partially correlated given other forecast errors through the common factors. Therefore, we impose a sparsity assumption on the precision matrix of the idiosyncratic errors, $\mathbf{\Theta}_\varepsilon$, which is obtained using the estimated residuals after removing the co-movements induced by the factors (see Barigozzi et al. (2018); Brownlees et al. (2018); Koike (2020)).

We use the weighted Graphical Lasso and nodewise regression as shrinkage techniques to estimate the precision matrix of residuals. Let $\mathbf{\Theta}_\varepsilon = \mathbf{\Sigma}_\varepsilon^{-1}$ and $\mathbf{\Theta}_f = \mathbf{\Sigma}_f^{-1}$ be the precision matrices of the idiosyncratic and common components respectively. Once the precision of the low-rank component is obtained, similarly to Fan et al. (2011), we use the Sherman-Morrison-Woodbury formula to estimate the precision of forecast errors:

$$\mathbf{\Theta} = \mathbf{\Theta}_\varepsilon - \mathbf{\Theta}_\varepsilon \mathbf{B}[\mathbf{\Theta}_f + \mathbf{B}'\mathbf{\Theta}_\varepsilon \mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_\varepsilon. \tag{4.1}$$

To obtain $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$, we use $\widehat{\mathbf{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\mathbf{f}}_t - \bar{\mathbf{f}})(\widehat{\mathbf{f}}_t - \bar{\mathbf{f}})'$, where $\widehat{\mathbf{f}}_t = \widehat{\mathbf{B}}'\mathbf{e}_t$. To get $\widehat{\mathbf{\Theta}}_\varepsilon$, we develop two approaches: the first uses the weighted GLASSO Algorithm 1, with the initial estimate of the covariance matrix of the idiosyncratic errors calculated as $\widehat{\mathbf{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\boldsymbol{\varepsilon}}_t - \bar{\boldsymbol{\varepsilon}})(\widehat{\boldsymbol{\varepsilon}}_t - \bar{\boldsymbol{\varepsilon}})'$, where $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$. The second uses nodewise regression and applies Algorithm 2 to $\widehat{\varepsilon}$. Once we estimate $\widehat{\mathbf{\Theta}}_f$ and $\widehat{\mathbf{\Theta}}_\varepsilon$, we can get $\widehat{\mathbf{\Theta}}$ using a sample analogue of (4.1). We call the proposed procedures *Factor Graphical Lasso* and *Factor nodewise regression* and summarize them in Algorithm 3 and Algorithm 4 respectively.

Algorithm 3 Factor Graphical Lasso (Factor GLASSO)

1: Estimate the residuals: $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$ using PCA.
Get $\widehat{\mathbf{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\boldsymbol{\varepsilon}}_t - \bar{\boldsymbol{\varepsilon}})(\widehat{\boldsymbol{\varepsilon}}_t - \bar{\boldsymbol{\varepsilon}})'$.
2: Estimate a sparse $\mathbf{\Theta}_\varepsilon$ using the weighted Graphical Lasso: initialize Algorithm 1 with $\mathbf{W} = \widehat{\mathbf{\Sigma}}_\varepsilon + \lambda\mathbf{I}$.
3: Estimate $\mathbf{\Theta}$ using the Sherman-Morrison-Woodbury formula in (4.1).

Algorithm 4 Factor nodewise regression Meinshausen and Bühlmann (2006) (Factor MB)

1: Estimate the residuals: $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$ using PCA.
Get $\widehat{\mathbf{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}(\widehat{\boldsymbol{\varepsilon}}_t - \bar{\boldsymbol{\varepsilon}})(\widehat{\boldsymbol{\varepsilon}}_t - \bar{\boldsymbol{\varepsilon}})'$.
2: Estimate a sparse $\mathbf{\Theta}_\varepsilon$ using nodewise regression: apply Algorithm 2 to $\widehat{\varepsilon}$.
3: Estimate $\mathbf{\Theta}$ using the Sherman-Morrison-Woodbury formula in (4.1).

Now we can use $\widehat{\mathbf{\Theta}}$ to estimate the forecast combination weights $\widehat{\mathbf{w}}$

$$\widehat{\mathbf{w}} = \frac{\widehat{\mathbf{\Theta}}\boldsymbol{\iota}_p}{\boldsymbol{\iota}_p'\widehat{\mathbf{\Theta}}\boldsymbol{\iota}_p}, \tag{4.2}$$

where $\widehat{\mathbf{\Theta}}$ is obtained from Algorithm 3 or Algorithm 4.

4.1 The Choice of the Tuning Parameters for FGM

Algorithms 3-4 require the tuning parameters λ (from Algorithm 1) and λ_j (from Algorithm 2) respectively. We now briefly comment on the choices for both tuning parameters.

To motivate the choice of the tuning parameter for GLASSO and Factor GLASSO, we first briefly discuss some of the existing options to motivate our choice of λ in (2.1) in simulations and the empirical application. Usually λ is selected from a grid of values $F_\lambda = (\lambda_{\min}, \dots, \lambda_{\max})$ which minimizes the score measuring the goodness-of-fit. Some popular examples include multifold cross-validation (CV), Stability Approach to Regularization Selection (STARS, Liu et al. (2010)), and the Extended Bayesian Information Criteria (EBIC, Foygel and Drton (2010)). Since we are interested in estimating a sparse high-dimensional precision matrix, we need to choose a method for selecting the tuning parameter which is consistent in high-dimensions. Meinshausen and Bühlmann (2010) suggest that CV performs poorly for high-dimensional data, it overfits (Liu et al. (2010)), and it does not consistently select models (Shao (1993)). Zhu and Cribben (2018) pointed out that the STARS is not computationally efficient. It is consistent under certain conditions, but suffers from the problem of overselection in estimating Gaussian graphical models. In contrast, EBIC is computationally efficient and is considered to be the state-of-the-art technique for choosing the tuning parameter for the undirected graphs. The score measuring the goodness of fit for EBIC can be written as:

$$\lambda_{\text{EBIC}} = \arg\min_{\lambda \in F_\lambda} \{-2l(\mathbf{\Theta}_\lambda) + \log(T)\text{df}(\mathbf{\Theta}_\lambda) + 4\text{df}(\mathbf{\Theta}_\lambda)\log(p)\eta\}, \tag{4.3}$$

where $\eta \in [0, 1]$, $\mathbf{\Theta}_\lambda$ is the precision matrix estimated for the tuning parameter $\lambda \in F_\lambda$, and the log-likelihood is $l(\mathbf{\Theta}_\lambda) = \log\det(\mathbf{\Theta}_\lambda) - \text{trace}(\mathbf{S}\mathbf{\Theta}_\lambda)$. For the estimation of graphical models, the degrees of freedom are usually defined as the number of unique non-zero elements in the estimated precision matrix, $\text{df}(\mathbf{\Theta}_\lambda) = \sum_{i \le j} I_{\mathbf{\Theta}_{\lambda,i,j} \neq 0}$. When $\eta = 0$, (4.3) reduces to the original BIC (Schwarz (1978)). Chen and Chen (2008) showed that when $\gamma = 1$, EBIC is consistent as long as the dimension p

does not grow exponentially with the sample size T. Hence, in our simulations and the empirical exercise we use EBIC with $\eta = 1$ for GLASSO and Factor GLASSO in Algorithms 1 and 3.

For Algorithms 2 and 4, we follow Callot et al. (2019) to choose λ_j in (2.2) by minimizing the generalized information criterion (GIC) developed by Fan and Tang (2013). Let $\left|\widehat{S}_j(\lambda_j)\right|$ denote the estimated number of nonzero parameters in the vector $\widehat{\boldsymbol{\gamma}}_j$:

$$\text{GIC}(\lambda_j) = \log(\|\mathbf{e}_j - \mathbf{E}_{-j}\widehat{\boldsymbol{\gamma}}_j\|_2^2/T) + \left|\widehat{S}_j(\lambda_j)\right| \frac{\log(p)}{T} \log(\log(T)). \tag{4.4}$$

5 Monte Carlo

We divide the simulation results into two subsections. In the first subsection we study the consistency of the Factor GLASSO and Factor MB for estimating precision matrix and the combination weights. In the second subsection we evaluate the out-of-sample forecasting performance of the Factor Graphical models from Algorithms 3-4 in terms of the mean-squared forecast error. We compare the performance of factor-based models with equal-weighted (EW) forecast combination, GLASSO and nodewise regression from Algorithms 1-2. All exercises use 100 Monte Carlo simulations.

5.1 Consistent Estimation of forecast combination weights based on FGM

We consider sparse Gaussian graphical models which may be fully specified by a precision matrix $\boldsymbol{\Theta}_0$. Therefore, the random sample is distributed as $\mathbf{e}_t = (e_{1t}, \ldots, e_{pt})' \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Theta}_0 = (\boldsymbol{\Sigma}_0)^{-1}$ for $t = 1, \ldots, T,\ j = 1, \ldots, p$. Let $\widehat{\boldsymbol{\Theta}}$ be the precision matrix estimator. We show consistency of the Factor GLASSO (Algorithm 3) and Factor MB (Algorithm 4), in (i) the operator norm, $\left|\left|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right|\right|\right|_2$, and (ii) l_1/l_1-matrix norm which is the maximum absolute column sum of the matrix , $\left|\left|\left|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\right|\right|\right|_1$, and (iii) in l_1-vector norm for the combination weights, $\|\widehat{\mathbf{w}} - \mathbf{w}\|_1$, where $\mathbf{w}$ is given by (3.7).

5.1.1 Simulation Design

The forecast errors are assumed to have the following structure:

$$\mathbf{f}_t = \phi_f \mathbf{f}_{t-1} + \boldsymbol{\zeta}_t \tag{5.1}$$

$$\underbrace{\mathbf{e}_t}_{p\times 1} = \mathbf{B} \underbrace{\mathbf{f}_t}_{q\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{5.2}$$

where $\mathbf{e}_t$ is a $p \times 1$ vector of forecast errors following $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{f}_t$ is a $q \times 1$ vector of factors, $\mathbf{B}$ is a $p \times q$ matrix of factor loadings, ϕ_f is an autoregressive parameter in the factors which is a scalar for simplicity, $\boldsymbol{\zeta}_t$ is a $q \times 1$ random vector with each component independently following $\mathcal{N}(0, \sigma_\zeta^2)$, $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ random vector following $\mathcal{N}(0, \mathbf{\Sigma}_\varepsilon)$, with sparse $\mathbf{\Theta}_\varepsilon$ that has a random graph structure described below. To create $\mathbf{B}$ in (5.1) we take the first q columns of an upper triangular matrix from a Cholesky decomposition of the $p \times p$ Toeplitz matrix parameterized by ρ: that is, $\mathbf{Q} = (\mathbf{Q})_{ij}$, where $(\mathbf{Q})_{ij} = \rho^{|i-j|}$, $i, j \in 1, \ldots, p$. We set $\rho = 0.2$, $\phi_f = 0.2$ and $\sigma_\zeta^2 = 1$. The specification in (5.1) leads to the low-rank plus sparse decomposition of the covariance matrix:

$$\mathbb{E}\left[\mathbf{e}_t \mathbf{e}_t'\right] = \mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_f \mathbf{B}' + \mathbf{\Sigma}_\varepsilon \tag{5.3}$$

When $\mathbf{\Sigma}_\varepsilon$ has a sparse inverse $\mathbf{\Theta}_\varepsilon$, it leads to the low-rank plus sparse decomposition of the precision matrix $\mathbf{\Theta}$, such that $\mathbf{\Theta}$ can be expressed as a function of the low-rank $\mathbf{\Theta}_f$ plus sparse $\mathbf{\Theta}_\varepsilon$.

We consider the following setup: let $p = T^\delta$, $\delta = 0.85$, $q = 2(\log(T))^{0.5}$ and $T = [2^\kappa]$, for $\kappa = 7, 7.5, 8, \ldots, 9.5$. Our setup allows the number of individual forecasts, p, and the number of common factors in the forecast errors, q, to increase with the sample size, T.

A sparse precision matrix of the idiosyncratic components $\mathbf{\Theta}_\varepsilon$ is constructed as follows: we first generate the adjacency matrix using a random graph structure. Define a $p \times p$ adjacency matrix $\mathbf{A}$ which is used to represent the structure of the graph:

$$\mathbf{A}^{ij} = \begin{cases} 1, & \text{for } i \neq j \text{ with probability } \pi, \\ 0, & \text{otherwise.} \end{cases} \tag{5.4}$$

Let $\mathbf{A}_{\varepsilon,ij}$ denote the i,j-th element of the adjacency matrix $\mathbf{A}_\varepsilon$. We set $\mathbf{A}_{\varepsilon,ij} = \mathbf{A}_{\varepsilon,ji} = 1$, for $i \neq j$ with probability π, and 0 otherwise. Such structure results in $s_T = p(p-1)\pi/2$ edges in the graph. To control sparsity, we set $\pi = 1/(pT^{0.8})$, which makes $s_T = \mathcal{O}(T^{0.05})$. The adjacency matrix has all diagonal elements equal to zero. Hence, to obtain a positive definite precision matrix we apply the procedure described in Zhao et al. (2012): using their notation, $\mathbf{\Theta}_\varepsilon = \mathbf{A} \cdot v + \mathbf{I}(|\tau| + 0.1 + u)$, where $u > 0$ is a positive number added to the diagonal of the precision matrix to control the magnitude of partial correlations, v controls the magnitude of partial correlations with u, and τ is the smallest eigenvalue of $\mathbf{A} \cdot v$. In our simulations we use $u = 0.1$ and $v = 0.3$.

5.1.2 Simulation Results

Figures 2-3 show the averaged (over Monte Carlo simulations) errors of the estimators of the precision matrix $\mathbf{\Theta}$ and the optimal combination weight versus the sample size T in the logarithmic scale (base 2). The estimate of the precision matrix of the EW forecast is obtained using the fact that equal weights imply diagonal covariance and precision matrices. To determine the values of the diagonal elements we use the shrinkage intensity coefficient calculated as the average of the eigenvalues of the sample covariance matrix of the forecast errors (see Ledoit and Wolf (2004b)). As evidenced by Figures 2-3, Factor GLASSO and Factor MB demonstrate superior performance over EW and non-factor based models (GLASSO and MB). Furthermore, our method achieves lower estimation error in the combination weights (3.9), which leads to lower risk of the combined forecast as shown in (3.8). Interestingly, even though the precision matrix estimated using Factor MB has faster convergence rate in $|||\cdot|||_2$ and $|||\cdot|||_1$ norms as compared to Factor GLASSO, the weights estimated using Factor GLASSO converge faster. Also, note that the precision matrix estimated using the EW method also shows good convergence properties. However, in terms of estimating the combination weight (and, as a corollary of (3.9), controlling risk), the performance of the EW does not exhibit convergence properties. This is in agreement with previously reported findings (see Claeskens et al. (2016); Smith and Wallis (2009) among others) that equal weights are not theoretically optimal, however, as demonstrated in Figure 4, the EW combination still leads to a relatively good performance in terms of MSFE.

5.2 Comparing Performance of forecast combinations based on FGM

We consider the standard forecasting model in the literature (e.g., Stock and Watson (2002)), which uses the factor structure of the high dimensional predictors.

5.2.1 Simulation Design

Suppose the data is generated from the following data generating process (DGP):

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{g}_t + \mathbf{v}_t, \tag{5.5}$$

$$\mathbf{g}_t = \phi\mathbf{g}_{t-1} + \boldsymbol{\xi}_t, \tag{5.6}$$

$$y_{t+1} = \mathbf{g}_t'\boldsymbol{\alpha} + \sum_{s=1}^{\infty}\theta_s\epsilon_{t+1-s} + \epsilon_{t+1}, \tag{5.7}$$

where y_{t+1} is a univariate series of our interest in forecasting, $\mathbf{x}_t$ is an $N \times 1$ vector of regressors (predictors), $\boldsymbol{\beta}$ is an $N \times 1$ parameter vector, $\mathbf{g}_t$ is an $r \times 1$ vector of factors, $\boldsymbol{\Lambda}$ is an $N \times r$ matrix of factor loadings, $\mathbf{v}_t$ is an $N \times 1$ random vector following $\mathcal{N}(0, \sigma_v^2)$, ϕ is an autoregressive parameter in the factors which is a scalar for simplicity, $\boldsymbol{\xi}_t$ is an $r \times 1$ random vector with each component independently following $\mathcal{N}(0, \sigma_\xi^2)$, ϵ_{t+1} is a random error following $\mathcal{N}(0, \sigma_\epsilon^2)$, and $\boldsymbol{\alpha}$ is an $r \times 1$ parameter vector which is drawn randomly from $\mathcal{N}(1, 1)$. We set $\sigma_\epsilon = 1$. The coefficients θ_s are set according to the rule

$$\theta_s = (1+s)^{c_1} c_2^s, \tag{5.8}$$

as in Hansen (2008). We set $c_1 \in \{0, 0.75\}$ and $c_2 \in \{0.6, 0.7, 0.8, 0.9\}$. We generate r factors using (5.6) with a grid of 10 different AR(1) coefficients ϕ equidistant between 0 and 0.9. To create $\boldsymbol{\Lambda}$ in (5.5) we take the first r rows of an upper triangular matrix from a Cholesky decomposition of the $N \times N$ Toeplitz matrix parameterized by ρ. We consider a grid of 10 different values of ρ equidistant between 0 and 0.9.

One-step ahead forecasts are estimated from the factor-augmented autoregressive (FAR) models of orders k, l, denoted as FAR(k, l):

$$\hat{y}_{t+1} = \hat{\mu} + \hat{\kappa}_1 \hat{g}_{1,t} + \cdots + \hat{\kappa}_k \hat{g}_{k,t} + \hat{\psi}_1 y_t + \cdots + \hat{\psi}_l y_{t+1-l}, \tag{5.9}$$

where the factors $(\hat{g}_{1,t}, \ldots, \hat{g}_{k,t})$ are estimated from equation (5.5). We consider the FAR models of various orders, with $k = 1, \ldots, K$ and $l = 1, \ldots, L$. We also consider the models without any lagged y or any factors. Therefore, the total number of forecasting models is $p \equiv (1+K) \times (1+L)$, which includes the forecasting models using naive average or no factors.

The total number of observations is T, and the number of observations in the regression period (the train sample) is set to be the first half of the sample, $t = 1, \ldots, m \equiv T/2$, to leave the second half of the sample, $t = m+1, \ldots, T$, for the out-of-sample evaluation (the test sample). We roll the estimation window over the test sample of the size $n \equiv T - m$, to update all the estimates in each point of time $t = 1, \ldots, m$. Recall that q denotes the number of factors in the forecast errors as in equation (3.1). We first examine the properties of the combined forecasts based on the Factor Graphical models when T and p vary and compare their performance with the combined forecasts based on the GLASSO, MB and EW forecasts.

5.2.2 Simulation Results

We analyzed various scenarios for the MSFE simulation exercise. For the first set of simulations we consider a low-dimensional setup to demonstrate the advantage of using FGM even when the number of forecasts, p, is small relative to the sample size, T: (1) in such scenario EW has an advantage since there are not many models to combine and assigning equal weights should produce satisfactory performance, and (2) non-factor based models have the advantage over the models that estimate factors due to the estimation errors. As a result, this framework with the low-dimensional setup is favorable to EW and non-factor based models. Figure 4 shows the MSFE for different sample sizes and fixed parameters: we report the results for two values of $c_1 \in \{0, 0.75\}$. As evidenced from Figure 4, the models that use the factor structure outperform EW combination and non-factor based counterparts for both values of c_1.

Figures 5-9 show the performance in terms of MSFE for different number of predictors N, different values of c_2, ϕ, ρ and q: Factor-based models (Factor GLASSO and Factor MB) outperform the equal-weighted forecast combination and the standard GLASSO and nodewise regression without any factor structure. As evidenced from the figures, these findings are robust to the changes in the model parameters. Importantly, Figure 9 shows the scenario when the true number of principal components, r, is equal to 5, whereas none of the forecasters use PCA for prediction: in this case including at least 2 common components of the forecasting errors reduces MSFE, such that Factor GLASSO and Factor MB outperform EW forecast combination. Based on Figures 4-9, we see that Factor GLASSO, in general, has lower MSFE than Factor MB. This finding is further supported by our empirical application in Section 6.

6 Application of FGM for Macroeconomic Forecasting

An empirical application to forecasting macroeconomic time series in big data environment highlights the advantage of both Factor Graphical models described in Algorithms 3-4 in comparison with the existing methods of forecast combination. We use a large monthly frequency macroeconomic database of McCracken and Ng (2016), who provide a comprehensive description of the dataset and 128 macroeconomic series. We consider the time period 1960:1-2020:07 with the total number of observations $T = 726$, the training sample consists of $m = 120$ observations,

and the test sample $n \equiv T - m - h + 1$, where h is the forecast horizon. We roll the estimation window over the test sample to update all the estimates in each point of time $t = m, \ldots, T-h$. We estimate h-step ahead forecasts from FAR(k, l) with $k = 0, 1, \ldots, K = 9$, and $l = 0, 1, \ldots, L = 11$. The total number of forecasting models is $p = 120$. The optimal number of factors in the forecast errors (denoted as q in equation (3.1)) is chosen using the standard data-driven method that uses the information criterion IC1 described in Bai and Ng (2002). We note that in the majority of the cases the optimal number of factors was estimated to be equal to 1.

Table 1 compares the performance of the Factor GLASSO and Factor MB with the competitors for predicting four representative macroeconomic indicators of the US economy: monthly industrial production (INDPROD), S&P500 composite index, civilian Unemployment Rate (UNRATE), and the Effective Federal Funds rate (FEDFUNDS) using 127 remaining macroeconomic series. Let $\{Y_t\}_{t=1}^T$ be the series of interest for forecasting. Similarly to Coulombe et al. (2020), for INDPROD and S&P500, we forecast the average growth rate (with logs):

$$y_{t+h}^{(h)} = \frac{1}{h} \ln(Y_{t+h}/Y_t).$$

For UNRATE we forecast the average change (without logs):

$$y_{t+h}^{(h)} = \frac{1}{h}(Y_{t+h}/Y_t).$$

And for FEDFUNDS we forecast the log of the series:

$$y_{t+h}^{(h)} = \ln(Y_{t+h}).$$

As evidenced from Table 1, our methods outperform EW, GLASSO and nodewise regression: accounting for the factor structure results in lower MSFE. Therefore, the FGM framework developed in this paper leads to the superior performance of the combined forecast as compared to EW model even when the models/experts do not contain a lot of unique information. Our empirical application demonstrates that this finding does not originate from the difference in the performance of EW vs graphical models: as evidenced from Table 1, the performance of GLASSO is worse than that of EW for the FEDFUNDS series, wheres Factor GLASSO outperforms EW. Therefore, the improvement in the combined forecast comes from the use of the factor structure in the forecast errors. Note that in contrast with EW and non-factor based methods, the performance of Factor GLASSO and Factor MB does not deteriorate significantly when the forecast horizon, h, increases.

7 Conclusions

This paper proposed a novel precision matrix estimator for the forecast combination when the experts are assumed to make common mistakes. We account for the factor structure in the forecast errors by decomposing the precision matrix into a low-rank and sparse components, where the latter is estimated using the Graphical Lasso or nodewise regression. The proposed algorithms are called the Factor Graphical Models (Factor GLASSO and Factor MB). The framework developed in this paper overcomes the challenge of obtaining the forecasts that contain unique information, which was shown to be necessary to achieve a "winning" forecast combination. Our simulations demonstrate the consistency of the developed procedure for estimating the precision matrix and optimal forecast combination weights. An empirical application to forecasting macroeconomic time series in big data environment highlights the advantage of the FGM approach in comparison with the existing methods of forecast combination. It would be interesting to apply our model for the analysis of some prominent forecasting competition strategies, such as M4 competition. We leave this exercise for the future research.

References

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, 36(1):197–200.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

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

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Barigozzi, M., Brownlees, C., and Lugosi, G. (2018). Power-law partial correlation network models. *Electronic Journal of Statistics*, 12(2):2905–2929.

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operations Research*, 20(4):451–468.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg.

Brownlees, C., Nualart, E., and Sun, Y. (2018). Realized networks. *Journal of Applied Econometrics*, 33(7):986–1006.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Callot, L., Caner, M., Önder, A. O., and Ulaşan, E. (2019). A nodewise regression approach to estimating large portfolios. *Journal of Business & Economic Statistics*, 0(0):1–12.

Callot, L. A. F., Kock, A. B., and Medeiros, M. C. (2017). Modeling and forecasting large realized covariance matrices and portfolio choice. *Journal of Applied Econometrics*, 32(1):140–158.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Chan, Y. L., Stock, J. H., and Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review*, 1(2):91–121.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583.

Connor, G. and Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: application of a new test methodology. *Journal of Financial Economics*, 21(2):255–289.

Coulombe, P. G., Leroux, M., Stevanovic, D., and Surprenant, S. (2020). How is machine learning useful for macroeconomic forecasting? *arXiv:2008.12477*.

Dai, C., Lu, K., and Xiu, D. (2019). Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data. *Journal of Econometrics*, 208(1):43–79.

Diebold, F. X. and Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, 35(4):1679–1691.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383–1414.

Fan, J., Wang, W., and Zhong, Y. (2019). Robust covariance estimation for approximate factor models. *Journal of Econometrics*, 208(1):5–22.

Fan, J., Xue, L., and Yao, J. (2017). Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292–306.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, 75(3):531–552.

Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS, pages 604–612, USA. Curran Associates Inc.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350.

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

Hautsch, N., Kyj, L. M., and Oomen, R. C. A. (2012). A blocking and regularization approach to high-dimensional realized covariance estimation. *Journal of Applied Econometrics*, 27(4):625–645.

Janková, J. and van de Geer, S. (2018). Inference in high-dimensional graphical models. *Handbook of Graphical Models*, Chapter 14, pages 325–351. CRC Press.

Johnstone, I. M. and Lu, A. Y. (2009). Sparse principal components analysis. *arXiv:0901.4392*.

Koike, Y. (2020). De-biased graphical lasso for high-frequency data. *Entropy*, 22(4):456.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Ledoit, O. and Wolf, M. (2004a). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, 139:360–384.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pages 1432–1440, USA. Curran Associates Inc.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Thomson, M. E., Pollock, A. C., Önkal, D., and Gönül, M. S. (2019). Combining forecasts: performance and coherence. *International Journal of Forecasting*, 35(2):474–484.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, Vol. 1, Chapter 4, pages 135–196. Elsevier.

van de Geer, S., Buhlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vandenberghe, L., Boyd, S., and Wu, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19(2):499–533.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The HUGE package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(1):1059–1062.

Zhu, Y. and Cribben, I. (2018). Sparse graphical models for functional connectivity networks: Best methods and the autocorrelation issue. *Brain Connectivity*, 8(3):139–165. PMID: 29634321.

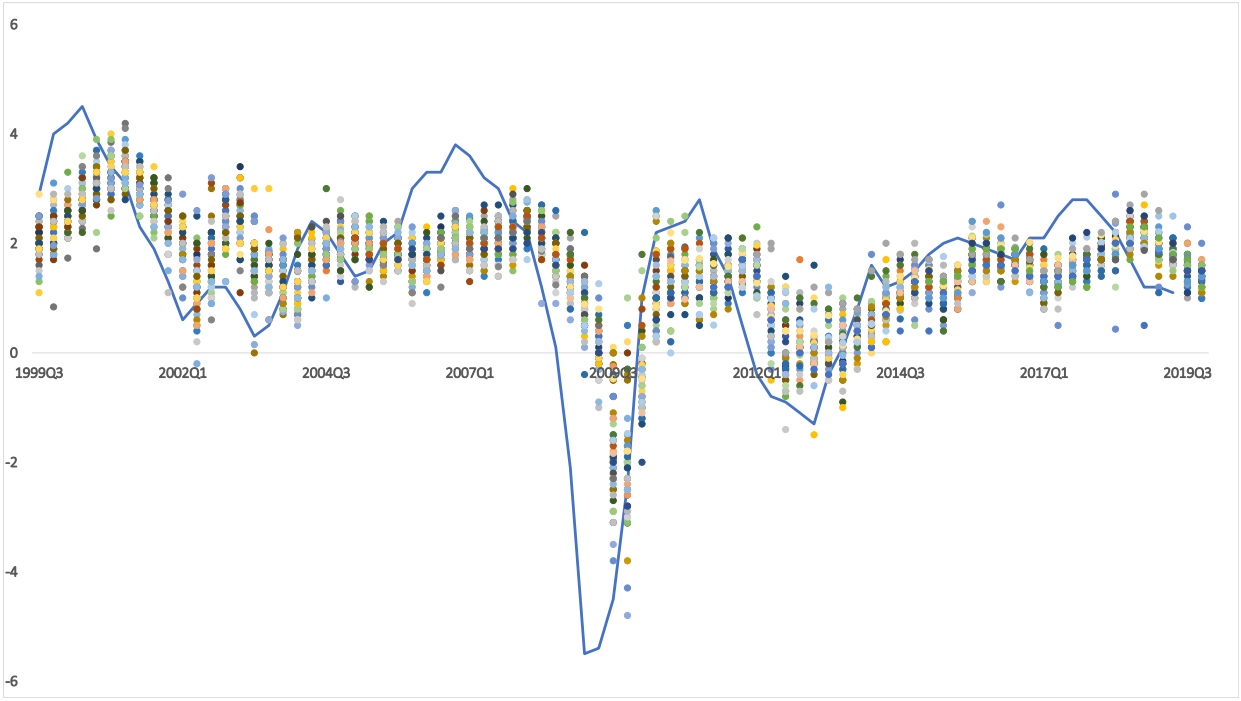


Figure 1: **The European Central Bank's (ECB) Survey of Professional Forecasters (SPF)**. Each circle denotes the forecast of each professional forecaster in the SPF for the quarterly 1-year-ahead forecasts of Euro-area real GDP growth, year-on-year percentage change. Actual series is the blue line. *Source: European Central Bank.*

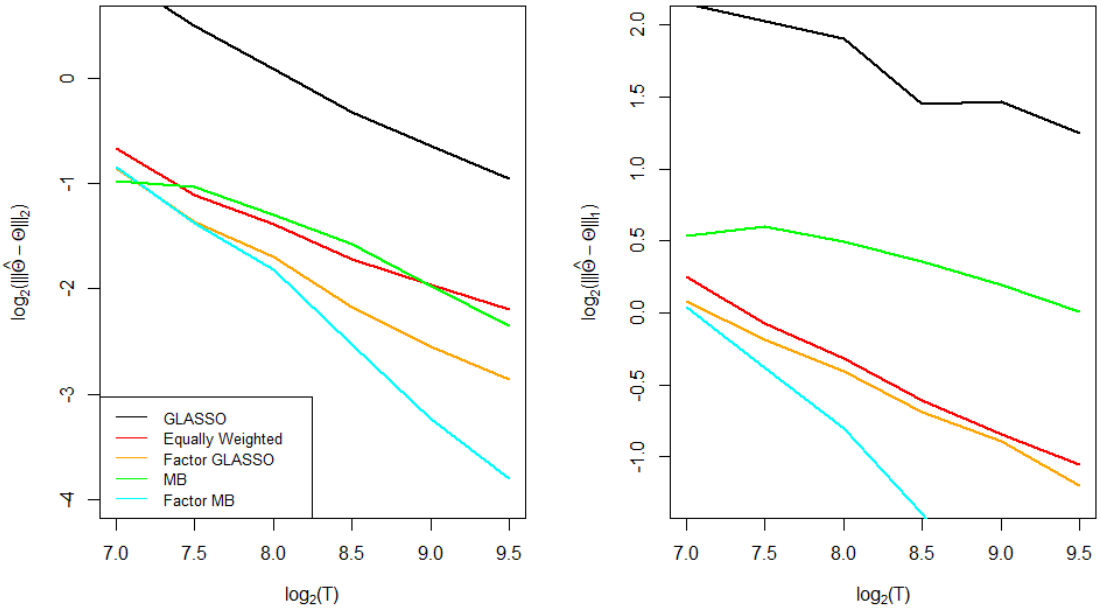


Figure 2: **Averaged errors of the estimators of $\mathbf{\Theta}$ on logarithmic scale (base 2):** $p = T^{0.85}$, $q = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.

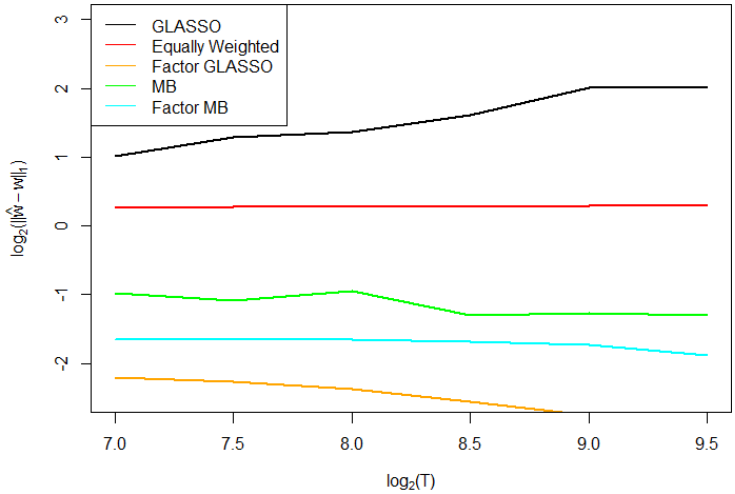
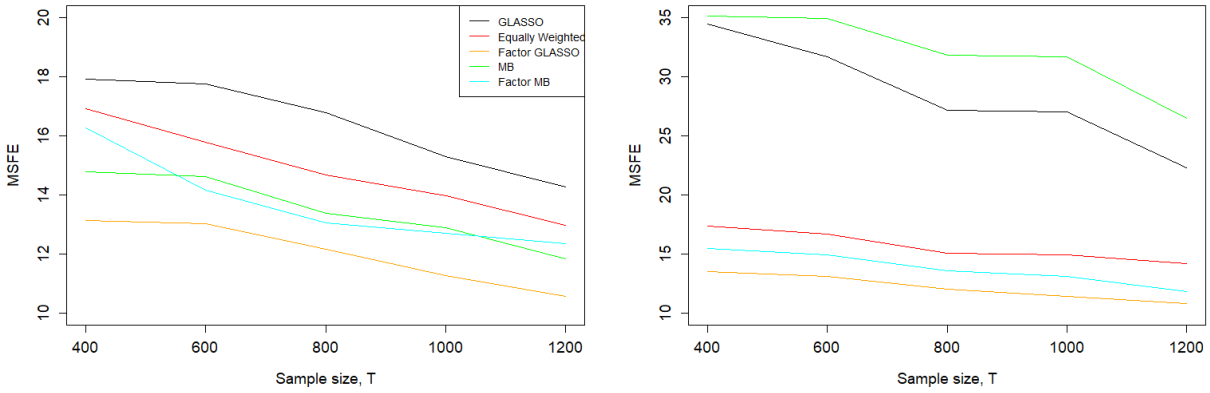


Figure 3: **Averaged errors of the estimator of w (base 2) on logarithmic scale:** $p = T^{0.85}$, $q = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.



(a) $c_1 = 0$. (b) $c_1 = 0.75$.

Figure 4: **Plots of the MSFE over the sample size T.** $c_1 \in \{0, 0.75\}$, $c_2 = 0.9$, $N = 100$, $r = 5$, $\sigma_\xi = 1$, $L = 7$, $K = 2$, $p = 24$, $q = 5$, $\rho = 0.9$, $\phi = 0.8$.

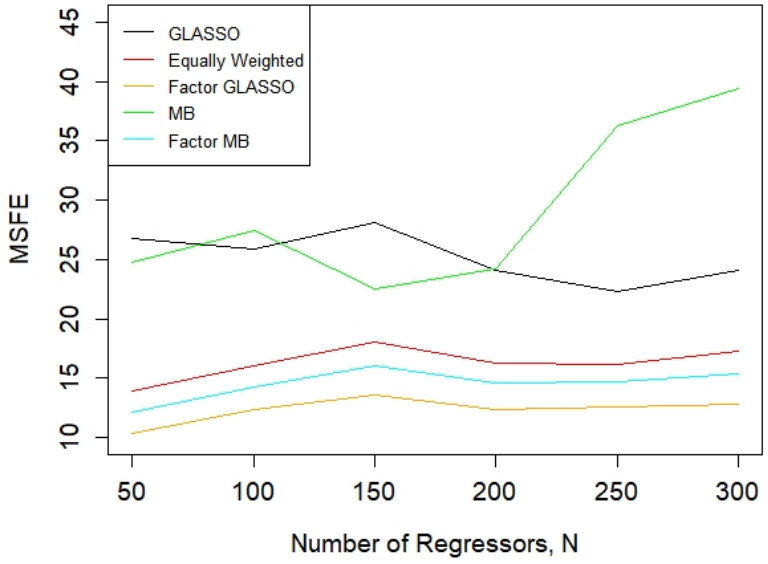


Figure 5: **Plots of the MSFE over the number of predictors N.** $c_1 = 0.75$, $c_2 = 0.9$, $T = 800$, $r = 5$, $\sigma_\xi = 1$, $L = 7$, $K = 2$, $p = 24$, $q = 5$, $\rho = 0.9$, $\phi = 0.8$.

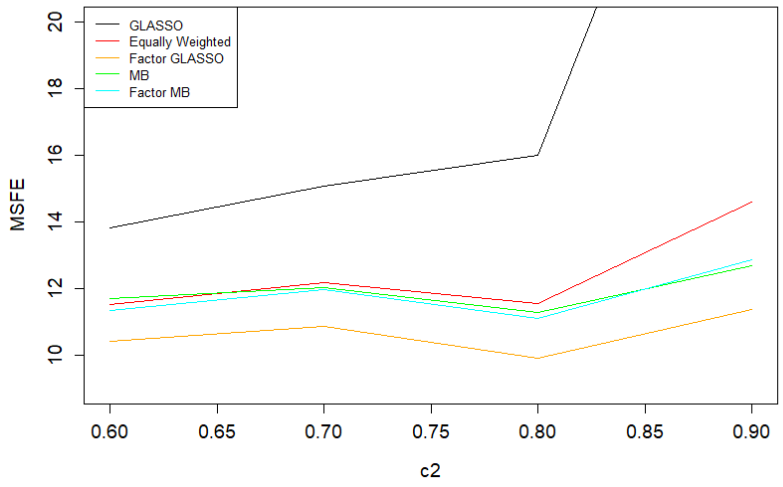


Figure 6: **Plots of the MSFE over the values of c_2.** $c_1 = 0.75,\ c_2 \in \{0.6, 0.7, 0.8, 0.9\},$ $T = 800,\ N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho = 0.9,\ \phi = 0.8.$

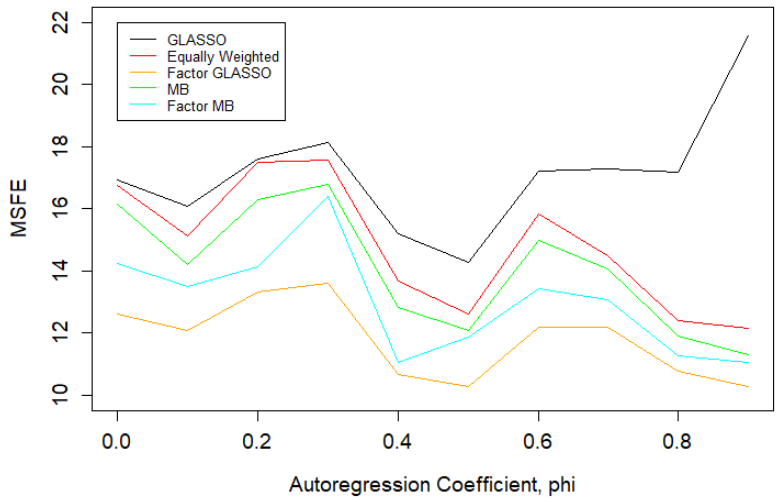


Figure 7: **Plots of the MSFE over the values of ϕ.** $c_1 = 0.75,\ c_2 = 0.8,\ T = 800,$ $N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho = 0.9,\ \phi \in \{0, 0.1, \ldots, 0.9\}.$

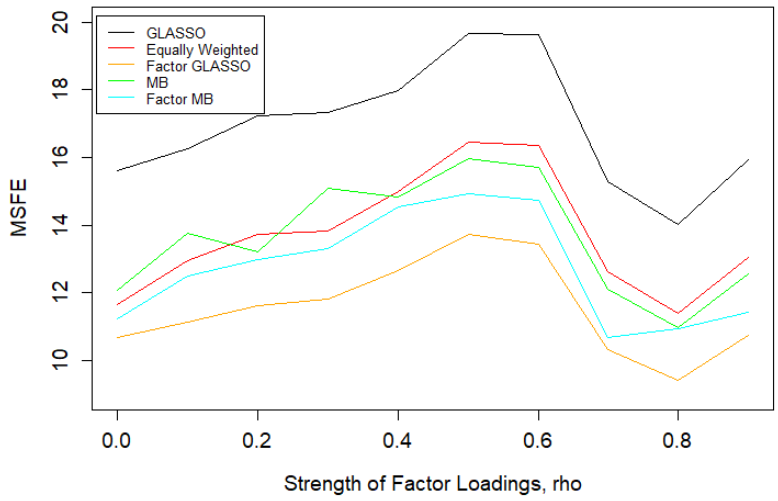


Figure 8: **Plots of the MSFE over the values of ρ.** $c_1 = 0.75,\ c_2 = 0.8,\ T = 800,$ $N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 7,\ K = 2,\ p = 24,\ q = 5,\ \rho \in \{0, 0.1, \ldots, 0.9\},\ \phi = 0.7.$

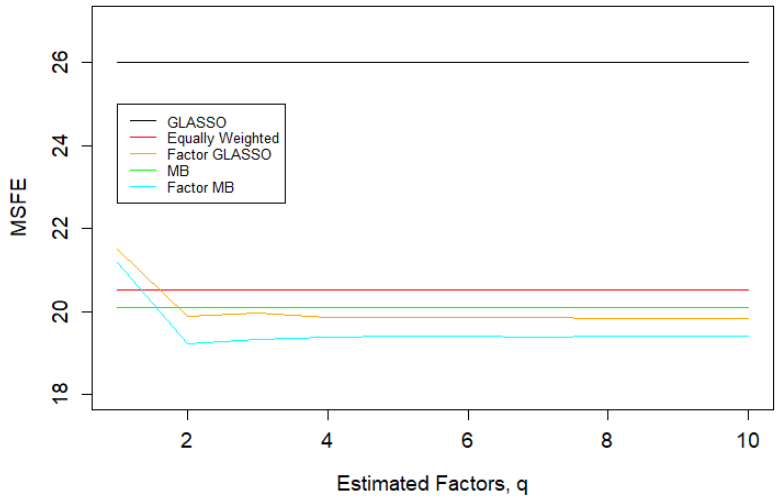


Figure 9: **Plots of the MSFE over the values of q.** $c_1 = 0.75,\ c_2 = 0.9,\ T = 800,$ $N = 100,\ r = 5,\ \sigma_\xi = 1,\ L = 12,\ K = 0,\ p = 13,\ q \in \{0, 1, \ldots, 10\},\ \rho = 0.9,\ \phi = 0.8.$

			INDPROD		
h	EW	GLASSO	Factor GLASSO	MB	Factor MB
1	2.77E-04	1.51E-04	1.24E-04	2.23E-04	1.28E-04
2	3.26E-04	1.79E-04	5.59E-05	1.61E-04	1.38E-04
3	1.55E-04	9.77E-05	3.81E-05	1.17E-04	6.54E-05
4	1.18E-04	7.60E-05	2.38E-05	1.03E-04	2.65E-05
			S&P500		
1	1.40E-03	1.39E-03	1.37E-03	1.34E-03	9.57E-03
2	1.71E-03	1.44E-03	8.95E-04	1.55E-03	1.01E-03
3	1.66E-03	1.34E-03	3.48E-04	1.43E-03	6.69E-04
4	1.27E-03	1.06E-03	3.95E-04	9.55E-04	7.91E-04
			UNRATE		
1	0.2531	0.0858	0.0109	0.0557	0.0107
2	0.3758	0.1334	0.0066	0.0448	0.0081
3	0.0743	0.0651	0.0066	0.0532	0.0051
4	2.1999	0.6871	0.1578	1.0973	0.2510
			FEDFUNDS		
1	0.0609	0.1813	0.0205	0.0424	0.0448
2	0.1426	1.2230	0.0288	0.0675	0.0416
3	0.2354	1.2710	0.0508	0.1217	0.1038
4	0.3702	1.4672	0.0592	0.2470	0.1962

Table 1: **Prediction of Monthly Macroeconomic Variables:** h indicates the forecast horizon, EW stands for the "Equal-Weighted" forecast, GLASSO and MB are the models that do not use the factor structure in the forecast errors. Factor GLASSO and Factor MB are our proposed Factor Graphical Models.