Combining Forecasts under Structural Breaks Using Graphical LASSO

Tae-Hwy Lee* and Ekaterina Seregina†

September 3, 2022

Abstract

In this paper we develop a novel method of combining many forecasts based on a machine learning algorithm called Graphical LASSO. We visualize forecast errors from different forecasters as a network of interacting entities and generalize network inference in the presence of common factor structure and structural breaks. First, we note that forecasters often use common information and hence make common mistakes, which makes the forecast errors exhibit common factor structures. We propose the Factor Graphical LASSO (Factor GLASSO), which separates common forecast errors from the idiosyncratic errors and exploits sparsity of the precision matrix of the latter. Second, since the network of experts changes over time as a response to unstable environments such as recessions, it is unreasonable to assume constant forecast combination weights. Hence, we propose Regime-Dependent Factor Graphical LASSO (RD-Factor GLASSO) and develop its scalable implementation using the Alternating Direction Method of Multipliers (ADMM) to estimate regime-dependent forecast combination weights. The empirical application to forecasting macroeconomic series using the data of the European Central Bank's Survey of Professional Forecasters (ECB SPF) demonstrates superior performance of a combined forecast using Factor GLASSO and RD-Factor GLASSO.

Keywords: Common Forecast Errors, Regime Dependent Forecast Combination, Sparse Precision Matrix of Idiosyncratic Errors, Structural Breaks.

JEL Classifications: C13, C38, C55

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

†Department of Economics, Colby College. Email: eseregin@colby.edu.

`Disclosure statement`: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors report there are no competing interests to declare.

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. 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".

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. For example, in the European Central Bank's Survey of Professional Forecasters (ECB SPF) of Euro-area real GDP growth, the forecasters tend to *jointly* understate or overstate GDP growth. Therefore, we stipulate that the forecast errors include common and idiosyncratic components, which allows the forecast errors to move together due to the common error component. Our paper provides a simple framework to learn from analyzing forecast errors: we separate unique errors from the common errors to improve the accuracy of the combined forecast.

Dating back to Bates and Granger (1969), the well-known expression for the optimal forecast combination weights requires an estimator of inverse covariance (precision) matrix. Precision matrix represents a network of interacting entities, such as corporations or genes. When the data is Gaussian, the sparsity in the precision matrix encodes the conditional independence graph - two variables are conditionally independent given the rest if

and only if the entry corresponding to these variables in the precision matrix is equal to zero. Graphical models are a powerful tool to directly estimate precision matrix, avoiding the step of obtaining an estimator of covariance matrix to be inverted. Prominent examples of graphical models include Graphical LASSO (Friedman et al. (2008)) and nodewise regression (Meinshausen and Bühlmann (2006)). Despite using different strategies for estimating precision matrix, all graphical models assume that precision matrix is sparse: many entries of precision matrix are zero, which is a necessary condition to consistently estimate inverse covariance. Our paper demonstrates that such assumption contradicts the stylized fact that experts tend to make common mistakes and hence the forecast errors move together through common factors. We show that graphical models fail to recover the entries of a nonsparse precision matrix under the factor structure.

This paper overcomes the aforementioned challenge and develops a new precision matrix estimator for the forecast errors under the approximate factor model with unobserved latent factors. We call our algorithm the *Factor Graphical LASSO*. We use a factor model to estimate an idiosyncratic component of the forecast errors, and then apply a Graphical LASSO for the estimation of the precision matrix of the idiosyncratic component.

At the same time, the network of experts changes over time, that is, the relationships between forecasts produced by different experts or models can change either smoothly or abruptly (e.g., as a response to an unexpected policy shock, or in the times of economic downturns). Such changes give rise to different regimes and it is important to account for changes in optimal forecast combination weights induced by structural breaks. This paper develops a unified framework to generalize network inference in the presence of structural breaks. We estimate regime-dependent precision matrix for forecast combination using both pre- and post-break data when forecast errors are driven by common factors. We call the

proposed algorithm *Regime-Dependent Factor Graphical LASSO* and develop its scalable implementation using the Alternating Direction Method of Multipliers (ADMM).

Our paper makes several contributions. First, we allow the forecast errors to be highly correlated due to the common component which is motivated by the stylized fact that the forecasters tend to jointly understate or overstate the predicted series of interest. Second, we develop a 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. We prove consistency of forecast combination weights and the Mean Squared Forecast Error (MSFE) estimated using Factor Graphical models. Third, to tackle changing relationships between forecasts produced by different experts or models as a response to unstable environments, we develop a unified framework to generalize network inference in the presence of structural breaks: we propose Regime-Dependent Factor Graphical LASSO (RD-Factor GLASSO) and develop its scalable implementation using ADMM to estimate regime-dependent forecast combination weights. Fourth, an empirical application to forecasting macroeconomic series using the data of the ECB SPF shows that incorporating (i) factor structure in the forecast errors together with (ii) sparsity in the precision matrix of the idiosyncratic components and (iii) regime-dependent combination weights improves the performance of a combined forecast over forecast combinations using equal weights.

The paper is structured as follows. Section 2 reviews Graphical LASSO. Section 3 studies the approximate factor model for the forecast errors. Section 4 introduces Factor Graphical LASSO and contains theoretical results on the consistency of the Factor GLASSO estimator. Section 5 introduces Regime-Dependent graphical model and discusses its implementation using ADMM. Section 6 validates theoretical results using simulations. Section 7 studies an

empirical application for macroeconomic time-series forecasting. Section 8 concludes.

Notation. For the convenience of the reader, we summarize the notation to be used throughout the paper. Let $\mathcal{S}_p$ denote the set of all $p \times p$ symmetric matrices. For any matrix $\mathbf{C}$, its (i,j)-th element is denoted as c_{ij}. 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 $\mathbf{U} \in \mathcal{S}_p$, let $\lambda_{\max}(\mathbf{U}) \equiv \lambda_1(\mathbf{U}) \geq \lambda_2(\mathbf{U}) \geq \ldots \geq \lambda_{\min}(\mathbf{U}) \equiv \lambda_p(\mathbf{U})$ be the eigenvalues of $\mathbf{U}$. Given a matrix $\mathbf{U} \in \mathbb{R}^{p\times p}$ and parameters $a, b \in [1,\infty)$, let $|||\mathbf{U}|||_{a,b} \equiv \max_{\|\mathbf{y}\|_a=1}\|\mathbf{U}\mathbf{y}\|_b$ denote the induced matrix-operator norm. The special cases are $|||\mathbf{U}|||_1 \equiv \max_{1\leq j\leq p}\sum_{i=1}^p |u_{ij}|$ for the ℓ_1/ℓ_1-operator norm; the operator norm (ℓ_2-matrix norm) $|||\mathbf{U}|||_2^2 \equiv \lambda_{\max}(\mathbf{U}\mathbf{U}')$ is equal to the maximal singular value of $\mathbf{U}$. Finally, $\|\mathbf{U}\|_{\max} \equiv \max_{i,j}|u_{ij}|$ denotes the element-wise maximum.

2 Graphical Models for Forecast Errors

This section briefly reviews a class of models, called graphical models, that search for the estimator of the precision matrix. 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.

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 $\mathbf{\Sigma}^{-1} \equiv \mathbf{\Theta}$ contains information about partial covariances between the variables. For instance, if θ_{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.

Let $\mathbf{W}$ be the estimate of $\mathbf{\Sigma}$. 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, which can be used as a choice for $\mathbf{W}$. Also, let $\widehat{\mathbf{\Gamma}}^2 \equiv \text{diag}(\mathbf{W})$ and its (i,j)-th element is denoted as $\widehat{\gamma}_{ij}$. We can write down truncated Gaussian log-likelihood (up to constants) $l(\mathbf{\Theta}) = \log\det(\mathbf{\Theta}) - \text{trace}(\mathbf{W}\mathbf{\Theta})$. When $\mathbf{W} = \mathbf{S}$, the maximum likelihood estimator of $\mathbf{\Theta}$ is $\widehat{\mathbf{\Theta}} = \mathbf{S}^{-1}$. The objective function associated with truncated Gaussian log-likelihood is also known as Bregman divergence and was shown to be applicable for non-Gaussian distributions (Ravikumar et al. (2011)).

In the high-dimensional settings it is necessary to regularize the precision matrix, which means that some edges will be zero. A natural way 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 penalized log-likelihood that weighs the variables by their scale:

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

over positive definite symmetric matrices, where $\tau \geq 0$ is a penalty parameter for the off-diagonal elements. We refer to the objective function in (2.1) as a "weighted penalized log-likelihood". The subscript τ in $\widehat{\mathbf{\Theta}}_\tau$ means that the solution of the optimization problem in (2.1) will depend upon the choice of the tuning parameter. In order to simplify notation, we will omit the subscript.

One of the 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). Define

the following partitions of $\mathbf{W}$, $\mathbf{S}$ and $\boldsymbol{\Theta}$:

$$\mathbf{W} = \begin{pmatrix} \underbrace{\mathbf{W}_{11}}_{(p-1)\times(p-1)} & \underbrace{\mathbf{w}_{12}}_{(p-1)\times 1} \\ \mathbf{w}_{12}' & w_{22} \end{pmatrix}, \mathbf{S} = \begin{pmatrix} \underbrace{\mathbf{S}_{11}}_{(p-1)\times(p-1)} & \underbrace{\mathbf{s}_{12}}_{(p-1)\times 1} \\ \mathbf{s}_{12}' & s_{22} \end{pmatrix}, \boldsymbol{\Theta} = \begin{pmatrix} \underbrace{\boldsymbol{\Theta}_{11}}_{(p-1)\times(p-1)} & \underbrace{\boldsymbol{\theta}_{12}}_{(p-1)\times 1} \\ \boldsymbol{\theta}_{12}' & \theta_{22} \end{pmatrix}. \tag{2.2}$$

Let $\boldsymbol{\beta} \equiv -\boldsymbol{\theta}_{12}/\theta_{22}$. The idea of GLASSO is to set $\mathbf{W} = \mathbf{S} + \tau\mathbf{I}$ in (2.1) and combine the gradient of (2.1) with the formula for partitioned inverses to obtain the following ℓ_1-regularized quadratic program

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \left\{ \frac{1}{2}\boldsymbol{\beta}'\mathbf{W}_{11}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{s}_{12} + \tau\widehat{\gamma}_{ii}\widehat{\gamma}_{jj}\|\boldsymbol{\beta}\|_1 \right\}. \tag{2.3}$$

As shown by Friedman et al. (2008), (2.3) can be viewed as a LASSO regression, where the LASSO estimates are functions of the inner products of $\mathbf{W}_{11}$ and s_{12}. Hence, (2.1) is equivalent to p coupled LASSO problems. Once we obtain $\widehat{\boldsymbol{\beta}}$, we can estimate the entries $\boldsymbol{\Theta}$ using the formula for partitioned inverses. The weighted GLASSO procedure is summarized in Algorithm 1.

Algorithm 1 Weighted Graphical LASSO

1: Initialize $\mathbf{W} = \mathbf{S} + \tau\mathbf{I}$, with $w_{ii} = s_{ii}$. The diagonal of $\mathbf{W}$ remains the same in what follows.

2: Estimate a sparse $\boldsymbol{\Theta}$ using the following weighted Graphical LASSO objective function from (2.1):

$$\widehat{\boldsymbol{\Theta}}_\tau = \arg\min_{\boldsymbol{\Theta}=\boldsymbol{\Theta}'} \operatorname{trace}(\mathbf{W}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta}) + \tau \sum_{i\neq j} \widehat{\gamma}_{ii}\widehat{\gamma}_{jj}|\theta_{ij}|,$$

over positive definite symmetric matrices.

3: 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:

$$\mathbf{W}_{11}\boldsymbol{\beta} - \mathbf{s}_{12} + \tau\widehat{\gamma}_{ii}\widehat{\gamma}_{jj} \cdot \operatorname{Sign}(\boldsymbol{\beta}) = \mathbf{0}.$$

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

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

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

As was shown in Friedman et al. (2008), the estimator produced by Algorithm 1 is guaranteed to be positive definite. Furthermore, Janková and van de Geer (2018) showed that Algorithm 1 is guaranteed to converge and produces consistent estimator of precision matrix under certain sparsity conditions.

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. We claim that the forecasters use the same set of public information to make forecasts and 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 ECB SPF 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 common 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, $\mathbf{f}_t$, and loadings, $\mathbf{B}$, are usually estimated by the principal component analysis (PCA), studied in Bai and Ng (2002); Stock

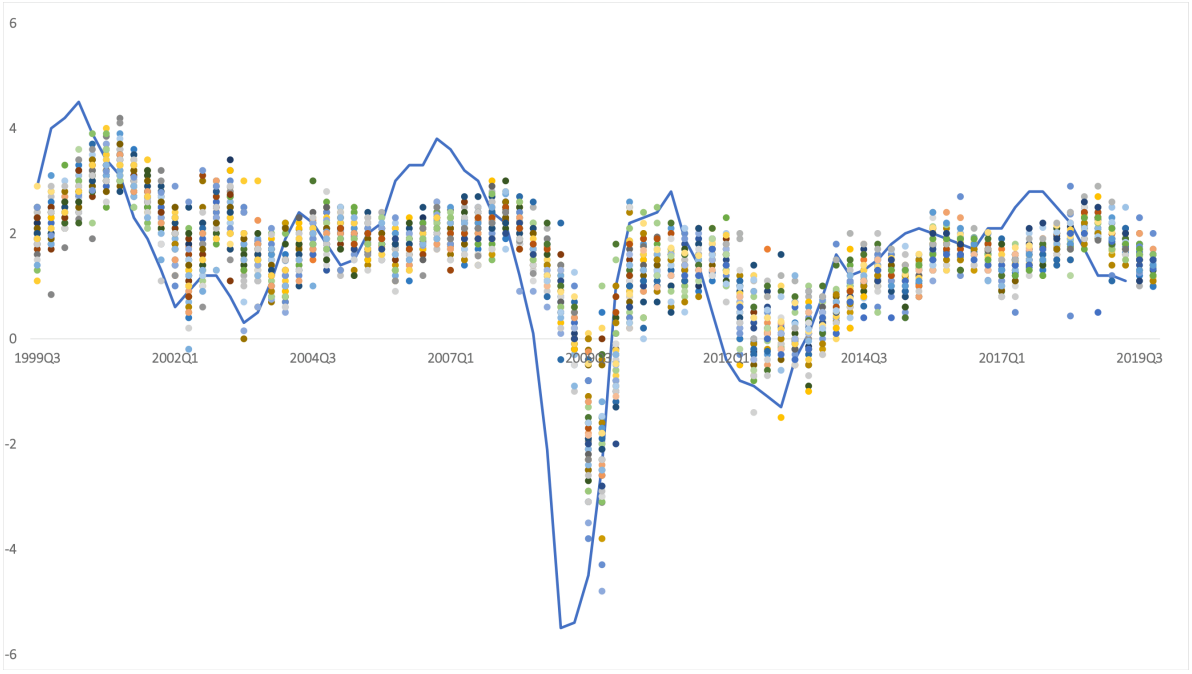

Figure 1: **The European Central Bank's (ECB) Survey of Professional Forecasters (SPF)**. Each circle denotes the forecast of each professional forecaster in the SPF for the quarterly 1-year-ahead forecasts of Euro-area real GDP growth, year-on-year percentage change. Actual series is the blue line.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

and Watson (2002). Strict factor structure assumes that the idiosyncratic forecast error terms, $\boldsymbol{\varepsilon}_t$, are uncorrelated with each other, whereas approximate factor structure allows correlation of the idiosyncratic components (Chamberlain and Rothschild (1983)).

We use the following notations: $\mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] = \boldsymbol{\Sigma}_\varepsilon$, $\mathbb{E}[\mathbf{f}_t\mathbf{f}_t'] = \boldsymbol{\Sigma}_f$, $\mathbb{E}[\mathbf{e}_t\mathbf{e}_t'] = \boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Sigma}_f\mathbf{B}' + \boldsymbol{\Sigma}_\varepsilon$, and $\mathbb{E}[\boldsymbol{\varepsilon}_t|\mathbf{f}_t] = 0$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Theta}_\varepsilon = \boldsymbol{\Sigma}_\varepsilon^{-1}$ and $\boldsymbol{\Theta}_f = \boldsymbol{\Sigma}_f^{-1}$ be the precision matrices of forecast errors, idiosyncratic and common components respectively. 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). Given a sample of the estimated residuals $\{\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t\}_{t=1}^{T}$ and the estimated factors $\{\widehat{\mathbf{f}}_t\}_{t=1}^{T}$, let $\widehat{\boldsymbol{\Sigma}}_\varepsilon = (1/T)\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$ and $\widehat{\boldsymbol{\Sigma}}_f = (1/T)\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$ be the sample counterparts of the covariance matrices.

Moving forward 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, $t = 1,\ldots,T$. 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 $\text{MSFE}(\mathbf{w},\boldsymbol{\Sigma}) = \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$. As shown in Bates and Granger (1969), the *optimal* forecast combination minimizes the MSFE of the combined forecast error:

$$\min_{\mathbf{w}} \text{MSFE} = \min_{\mathbf{w}} \mathbb{E}\left[\mathbf{w}'\mathbf{e}_t\mathbf{e}_t'\mathbf{w}\right] = \min_{\mathbf{w}} \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}, \ \text{s.t. } \mathbf{w}'\boldsymbol{\iota}_p = 1, \tag{3.6}$$

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), 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. Define $a = \boldsymbol{\iota}_p'\boldsymbol{\Theta}\boldsymbol{\iota}_p/p$ and $\widehat{a} = \boldsymbol{\iota}_p'\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p/p$. We can write

$$\left|\frac{\text{MSFE}(\widehat{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}})}{\text{MSFE}(\mathbf{w}, \boldsymbol{\Sigma})} - 1\right| = \left|\frac{\hat{a}^{-1}}{a^{-1}} - 1\right| = \frac{|a - \hat{a}|}{|\hat{a}|}, \tag{3.8}$$

and

$$\begin{aligned} \|\widehat{\mathbf{w}} - \mathbf{w}\|_1 &= \frac{\left[(a\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p) - (a\boldsymbol{\Theta}\boldsymbol{\iota}_p) + (a\boldsymbol{\Theta}\boldsymbol{\iota}_p) - (\hat{a}\boldsymbol{\Theta}\boldsymbol{\iota}_p)\right]/p}{(\hat{a}a)} \\ &\leq \frac{a\frac{\left\|(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1}{p} + |a - \widehat{a}|\frac{\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1}{p}}{|\widehat{a}|a}. \end{aligned} \tag{3.9}$$

Therefore, in order to control the estimation uncertainty in the MSFE and combination weights, one needs to obtain a consistent estimator of the precision matrix $\boldsymbol{\Theta}$. More details are discussed in section 4 and Theorem 1.

4 Factor Graphical LASSO for Forecast Errors

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. Hence, instead of imposing sparsity assumption on the precision of forecast errors, $\mathbf{\Theta}$, we require sparsity of the precision matrix of the idiosyncratic errors, $\mathbf{\Theta}_\varepsilon$. The latter is obtained using the estimated residuals after removing the co-movements induced by the factors (see Brownlees et al. (2018); Koike (2020)). Naturally, once we condition on the common components, it is sensible to assume that many remaining partial correlations of $\boldsymbol{\varepsilon}_t$ will be negligible and thus $\mathbf{\Theta}_\varepsilon$ is sparse.

We use the weighted Graphical LASSO as a shrinkage technique to estimate the precision matrix of residuals. Once the precision of the low-rank component is obtained, 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 $\widehat{\mathbf{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathbf{f}}_t \widehat{\mathbf{f}}_t'$. To get $\widehat{\mathbf{\Theta}}_\varepsilon$, we use the weighted GLASSO Algorithm 1, with the initial estimate of the covariance matrix of the idiosyncratic errors calculated as $\widehat{\mathbf{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\varepsilon}}_t \widehat{\boldsymbol{\varepsilon}}_t'$, where $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$. 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 procedure *Factor Graphical LASSO* and summarize it in Algorithm 2.

Algorithm 2 Factor Graphical LASSO (Factor GLASSO)

1: Estimate factors, $\widehat{\mathbf{f}}_t$, and factor loadings, $\widehat{\mathbf{B}}$, using PCA. Obtain $\widehat{\mathbf{\Sigma}}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{f}}_t\widehat{\mathbf{f}}_t'$, $\widehat{\mathbf{\Theta}}_f = \widehat{\mathbf{\Sigma}}_f^{-1}$, $\widehat{\boldsymbol{\varepsilon}}_t = \mathbf{e}_t - \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t$, and $\widehat{\mathbf{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\varepsilon}}_t\widehat{\boldsymbol{\varepsilon}}_t'$.

2: Estimate a sparse $\mathbf{\Theta}_\varepsilon$ using the weighted Graphical LASSO in (2.1) initialized with $\mathbf{W}_\varepsilon = \widehat{\mathbf{\Sigma}}_\varepsilon + \tau\mathbf{I}$:

$$\widehat{\mathbf{\Theta}}_{\varepsilon,\tau} = \arg\min_{\mathbf{\Theta}_\varepsilon = \mathbf{\Theta}_\varepsilon'} \operatorname{trace}(\mathbf{W}_\varepsilon\mathbf{\Theta}_\varepsilon) - \log\det(\mathbf{\Theta}_\varepsilon) + \tau\sum_{i\neq j}\widehat{\gamma}_{\varepsilon,ii}\widehat{\gamma}_{\varepsilon,jj}|\theta_{\varepsilon,ij}|. \tag{4.2}$$

where $\widehat{\gamma}_{\varepsilon,ii}$ is the (i,i)-th element of $\widehat{\mathbf{\Gamma}}_\varepsilon^2 \equiv \operatorname{diag}(\mathbf{W}_\varepsilon)$.

3: Use $\widehat{\mathbf{\Theta}}_f$ from Step 1 and $\widehat{\mathbf{\Theta}}_\varepsilon$ from Step 2 to estimate $\mathbf{\Theta}$ using the sample counterpart of the Sherman-Morrison-Woodbury formula in (4.1):

$$\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}_\varepsilon - \widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}[\widehat{\mathbf{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_\varepsilon. \tag{4.3}$$

Let $\widehat{\mathbf{\Theta}}_{\varepsilon,\tau}$ be the solution to (4.2) for a fixed τ. To choose the optimal shrinkage intensity coefficient, we minimize the following Bayesian Information Criterion (BIC) using grid search:

$$\text{BIC}(\tau) \equiv T\left[\operatorname{trace}(\widehat{\mathbf{\Theta}}_{\varepsilon,\tau}\widehat{\mathbf{\Sigma}}_\varepsilon) - \log\det(\widehat{\mathbf{\Theta}}_{\varepsilon,\tau})\right] + (\log T)\sum_{i\leq j}\mathbb{1}\left[\widehat{\theta}_{\varepsilon,\tau,ij}\neq 0\right]. \tag{4.4}$$

The grid $\mathcal{G} \equiv \{\tau_1,\ldots,\tau_M\}$ is constructed as follows: the maximum value in the grid, τ_M, is set to be the smallest value for which all the off-diagonal entries of $\widehat{\mathbf{\Theta}}_{\varepsilon,\tau_M}$ are zero, that is, the maximum modulus of the off-diagonal entries of $\widehat{\mathbf{\Sigma}}_\varepsilon$. The smallest value of the grid, $\tau_1 \in \mathcal{G}$, is determined as $\tau_1 \equiv \vartheta\tau_M$ for a constant $0 < \vartheta < 1$. The remaining grid values $\tau_1,\ldots,\tau_M$ are constructed in the ascending order from τ_1 to τ_M on the log scale:

$$\tau_i = \exp\Big(\log(\tau_1) + \frac{i-1}{M-1}\log(\tau_M/\tau_1)\Big), \quad i = 2,\ldots,M-1.$$

We use $\vartheta = \sqrt{\log p/T} + 1/\sqrt{p}$ (motivated by the convergence rate from Theorem 1) and

$M = 10$ in the simulations and the empirical exercise.

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

$$\widehat{\mathbf{w}} = \frac{\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p}{\boldsymbol{\iota}_p'\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p}, \tag{4.5}$$

where $\widehat{\boldsymbol{\Theta}}$ is obtained from Algorithm 2. This approach allows to extract the benefits of modeling common movements in forecast errors, captured by a factor model, and the benefits of using many competing forecasting models that give rise to a high-dimensional precision matrix, captured by a graphical model.

4.1 Asymptotic Analysis of the Estimator

We first introduce some terminology and notations. Let $A \in \mathcal{S}_p$. Define the following set for $j = 1, \ldots, p$:

$$D_j(A) \equiv \{i : A_{ij} \neq 0,\ i \neq j\}, \quad d_j(A) \equiv \text{card}(D_j(A)), \quad d(A) \equiv \max_{j=1,\ldots,p} d_j(A), \tag{4.6}$$

where $d_j(A)$ is the number of edges adjacent to the vertex j (i.e., the *degree* of vertex j), and $d(A)$ measures the maximum vertex degree. Define $S(A) \equiv \bigcup_{j=1}^{p} D_j(A)$ to be the overall off-diagonal sparsity pattern, and $s(A) \equiv \sum_{j=1}^{p} d_j(A)$ is the overall number of edges contained in the graph. Note that $\text{card}(S(A)) \leq s(A)$: when $s(A) = p(p-1)/2$ this would give a fully connected graph. We now list the assumptions on the model (3.1):

(A.1) (Spiked covariance model) Assume that (i) As $p \to \infty$, $\lambda_1(\boldsymbol{\Sigma}) > \lambda_2(\boldsymbol{\Sigma}) > \ldots > \lambda_q(\boldsymbol{\Sigma}) \gg \lambda_{q+1}(\boldsymbol{\Sigma}) \geq \ldots \geq \lambda_p(\boldsymbol{\Sigma}) > 0$, where $\lambda_j(\boldsymbol{\Sigma}) = \mathcal{O}(p)$ for $j \leq q$, while the non-spiked eigenvalues are bounded, that is, $c_0 \leq \lambda_j(\boldsymbol{\Sigma}) \leq C_0$, $j > q$ for constants

$c_0, C_0 > 0$.

And assume that (ii) $\boldsymbol{\iota}_p'\boldsymbol{\Theta}\boldsymbol{\iota}_p/p \geq c > 0$, where $c > 0$ is a positive constant.

(A.2) (Pervasive factors) There exists a positive definite $q \times q$ matrix $\breve{\mathbf{B}}$ such that

$$\left\|\left\|p^{-1}\mathbf{B}'\mathbf{B} - \breve{\mathbf{B}}\right\|\right\|_2 \to 0 \text{ and } \lambda_{\min}(\breve{\mathbf{B}})^{-1} = \mathcal{O}(1) \text{ as } p \to \infty.$$

We also impose strong mixing condition. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the σ-algebras that are generated by $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) : t \leq 0\}$ and $\{(\mathbf{f}_t, \boldsymbol{\varepsilon}_t) : t \geq T\}$ respectively. Define the mixing coefficient

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\Pr A \Pr B - \Pr AB|. \tag{4.7}$$

(A.3) (Strong mixing) There exists $r_3 > 0$ such that $3r_1^{-1} + 1.5r_2^{-1} + 3r_3^{-1} > 1$, and $C > 0$ satisfying, for all $T \in \mathbb{Z}^+$, $\alpha(T) \leq \exp(-CT^{r_3})$.

Assumption **(A.1)** divides the eigenvalues into the diverging and bounded ones. This assumption is satisfied by the factor model with pervasive factors, which is stated in Assumption **(A.2)**. We say that a factor is pervasive in the sense that it has non-negligible effect on a non-vanishing proportion of individual time-series. Part (ii) of Assumption **(A.1)** is needed for consistent estimation of the optimal forecast combination weights. Assumptions **(A.1)**-**(A.2)** are crucial for estimating a high-dimensional factor model: they ensure that the space spanned by the principal components in the population level $\boldsymbol{\Sigma}$ is close to the space spanned by the columns of the factor loading matrix $\mathbf{B}$. Assumption **(A.3)** is a technical condition which is needed to consistently estimate the factors and loadings. Note that Assumptions **(A.1)**(i), **(A.2)**, and **(A.3)** are standard assumptions and are used in Fan et al. (2013).

Let $\boldsymbol{\Lambda}_q = \text{diag}(\lambda_1, \ldots, \lambda_q)$ be a matrix of q leading eigenvalues of $\boldsymbol{\Sigma}$, and $\mathbf{V}_q = (\mathbf{v}_1, \ldots, \mathbf{v}_q)$ is a $p \times q$ matrix of their corresponding leading eigenvectors. Define $\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Lambda}}_q, \widehat{\mathbf{V}}_q$ to be the

estimators of $\mathbf{\Sigma}, \mathbf{\Lambda}_q, \mathbf{V}_q$. We further let $\widehat{\mathbf{\Lambda}}_q = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_q)$ and $\widehat{\mathbf{V}}_q = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_q)$ to be constructed by the first q leading empirical eigenvalues and the corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}\widehat{\mathbf{B}}' = \widehat{\mathbf{V}}_q\widehat{\mathbf{\Lambda}}_q\widehat{\mathbf{V}}_q'$. Similarly to Fan et al. (2018), we require the following bounds on the componentwise maximums of the estimators:

(B.1) $\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/T})$,

(B.2) $\left\|(\widehat{\mathbf{\Lambda}}_q - \mathbf{\Lambda}_q)\mathbf{\Lambda}_q^{-1}\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/T})$,

(B.3) $\left\|\widehat{\mathbf{V}}_q - \mathbf{V}_q\right\|_{\max} = \mathcal{O}_P(\sqrt{\log p/(Tp)})$.

Assumptions **(B.1)**-**(B.3)** are needed in order to ensure that the first q principal components are approximately the same as the columns of the factor loadings. The estimator $\widehat{\mathbf{\Sigma}}$ can be thought of as any "pilot" estimator that satisfies **(B.1)**. For sub-Gaussian distributions, sample covariance matrix, its eigenvectors and eigenvalues satisfy **(B.1)**-**(B.3)**.

In addition, the following structural assumptions on the model are imposed:

(C.1) $\|\mathbf{\Sigma}\|_{\max} = \mathcal{O}(1)$ and $\|\mathbf{B}\|_{\max} = \mathcal{O}(1)$.

Note that Assumptions **(B.1)**-**(B.3)** and **(C.1)** are standard assumptions and are used in Fan et al. (2018).

To study the properties of the combination weights in (4.5) and MSFE, we first need to establish the convergence properties of precision matrix produced by Algorithm 2. Let $\omega_T \equiv \sqrt{\log p/T} + 1/\sqrt{p}$. Also, let $s(\mathbf{\Theta}_\varepsilon) = \mathcal{O}_P(s_T)$ for some sequence $s_T \in (0, \infty)$ and $d(\mathbf{\Theta}_\varepsilon) = \mathcal{O}_P(d_T)$ for some sequence $d_T \in (0, \infty)$. The deterministic sequences s_T and d_T will control the sparsity $\mathbf{\Theta}_\varepsilon$ for Factor GLASSO. Note that d_T can be smaller than or equal to s_T.

Let ϱ_{1T} be a sequence of positive-valued random variables such that $\varrho_{1T}^{-1}\omega_T \xrightarrow{P} 0$ and $\varrho_{1T} d_T s_T \xrightarrow{P} 0$, with $\tau \asymp \omega_T$ (where τ is the tuning parameter for the Factor GLASSO in (4.2)). Lee and Seregina (2020) show that under the Assumptions **(A.1)**-**(A.3)**, **(B.1)**-**(B.3)** and **(C.1)**, $\left\|\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|\right\|_1 = \mathcal{O}_P(\varrho_{1T} d_T s_T) = o_P(1)$ and $\left\|\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right\|\right\|_2 = \mathcal{O}_P(\varrho_{1T} s_T) = o_P(1)$ for Factor GLASSO.

Having established the convergence rates for precision matrix, we now study the properties of the combination weights and the resulted MSFE.

Theorem 1. *Assume* ***(A.1)***-***(A.3)***, ***(B.1)***-***(B.3)***, *and* ***(C.1)*** *hold.*

(i) If $\varrho_{1T} d_T^2 s_T \xrightarrow{P} 0$, *Algorithm 2 consistently estimates forecast combination weights in (4.5):* $\|\widehat{\mathbf{w}} - \mathbf{w}\|_1 = \mathcal{O}_P\left(\varrho_{1T} d_T^2 s_T\right)$.

(ii) If $\varrho_{1T} d_T s_T \xrightarrow{P} 0$, *Algorithm 2 consistently estimates MSFE*$(\mathbf{w}, \mathbf{\Sigma})$*:* $\left|\frac{MSFE(\widehat{\mathbf{w}}, \widehat{\mathbf{\Sigma}})}{MSFE(\mathbf{w}, \mathbf{\Sigma})} - 1\right| = \mathcal{O}_P(\varrho_{1T} d_T s_T)$.

The proof of Theorem 1 can be found in Appendix A. Note that the rates of convergence for MSFE and precision matrix $\mathbf{\Theta}$ are the same and both are faster than the combination weight rates. In contrast to the classical graphical model in Algorithm 1, the convergence properties of which were examined by Janková and van de Geer (2018) among others, the rates in Theorem 1 depend on the sparsity of $\mathbf{\Theta}_\varepsilon$ rather than of $\mathbf{\Theta}$. This means that instead of assuming that many partial correlations of forecast errors $\mathbf{e}_t$ are negligible, which is not realistic under the factor structure, we impose a milder restriction requiring many partial correlations of $\boldsymbol{\varepsilon}_t$ to be negligible once the common components have been taken into account.

5 RD-Factor GLASSO for Forecast Errors

There are two streams of literature that study time-varying networks. The first one models dynamics in the precision matrix locally. Zhou et al. (2010) develop a nonparametric method for estimating time-varying graphical structure for multivariate Gaussian distributions using an ℓ_1-penalized log-likelihood. They find out that if the covariances change smoothly over time, the covariance matrix can be estimated well in terms of predictive risk even in high-dimensional problems. Lu et al. (2015) introduce nonparanormal graphical models that allow to model high-dimensional heavy-tailed systems and the evolution of their network structure. They show that the estimator consistently estimates the latent inverse Pearson correlation matrix. The second stream of literature allows the network to vary with time by introducing two different frequencies. Hallac et al. (2017) study time-varying Graphical LASSO with smoothing evolutionary penalty.

We augment the framework in Section 4 to account for regime switching by modeling the change in precision matrix due to N known structural breaks. Define $n_i \equiv t_i - t_{i-1}$ to be the sample between the i-th and $(i-1)$-th break points, where $i = 1, \ldots, N$, $\sum_{i=1}^{N} n_i = T$, $N \leq T$.

We propose that the dynamics of the system evolves through the precision matrix of the idiosyncratic component. Let $\mathbf{\Sigma}_{\varepsilon,i}$ and $\mathbf{\Sigma}_i$ be covariance matrices of idiosyncratic part and forecast errors in regime i. Define the corresponding precision matrices to be $\mathbf{\Theta}_{\varepsilon,i} \equiv \mathbf{\Sigma}_{\varepsilon,i}^{-1}$ and $\mathbf{\Theta}_i \equiv \mathbf{\Sigma}_i^{-1}$. We assume $\mathbf{\Sigma}_{f_i} = \mathbf{\Sigma}_f$ for all regimes i for simplicity but it can be generalized.

Let $\widehat{\mathbf{\Sigma}}_{\varepsilon,i} = \frac{1}{n_i} \sum_{k=1}^{n_i} \widehat{\boldsymbol{\varepsilon}}_{i,k} \widehat{\boldsymbol{\varepsilon}}_{i,k}'$. To model dynamics in $\{\mathbf{\Theta}_{\varepsilon,i}\}_{i=1}^{N}$ we use the following opti-

mization problem:

$$\{\widehat{\mathbf{\Theta}}_{\varepsilon,i}\}_{i=1}^{N} = \arg\min_{\{\mathbf{\Theta}_{\varepsilon,i}\}_{i=1}^{N}\succ 0} \sum_{i=1}^{N} n_i \Big[\operatorname{trace}\Big(\widehat{\mathbf{\Sigma}}_{\varepsilon,i}\mathbf{\Theta}_{\varepsilon,i}\Big) - \log\det\mathbf{\Theta}_{\varepsilon,i}\Big] + \alpha\|\mathbf{\Theta}_{\varepsilon,i}\|_{\text{od},1} + \beta\sum_{i=2}^{N}\psi(\mathbf{\Theta}_{\varepsilon,i} - \mathbf{\Theta}_{\varepsilon,i-1}). \tag{5.1}$$

where the penalty for the off-diagonal (od) elements is $\|\mathbf{\Theta}_{\varepsilon,i}\|_{\text{od},1} = \sum_{l\neq q}\widehat{\gamma}_{\varepsilon,ll,i}\widehat{\gamma}_{\varepsilon,qq,i}|\theta_{\varepsilon,lq,i}|$, $\widehat{\gamma}_{\varepsilon,ll,i}$ is the (l,l)-th element of $\widehat{\mathbf{\Gamma}}^2_{\varepsilon,i} \equiv \text{diag}(\widehat{\mathbf{\Sigma}}_{\varepsilon,i})$ and $\theta_{\varepsilon,lq,i}$ is the lq-th element of matrix $\mathbf{\Theta}_{\varepsilon,i}$. Figure 2 visualizes dynamics of the precision matrix.

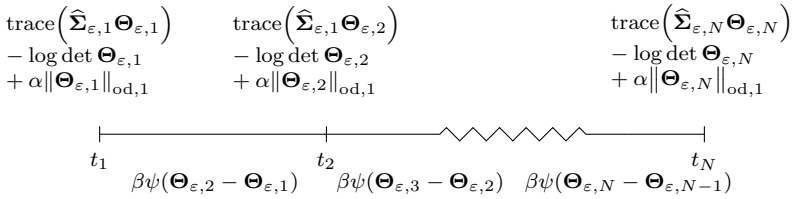

Figure 2: **Change of precision matrix over time:** β is the penalty that enforces temporal consistency and ψ is a convex penalty function.

The optimization problem in (5.1) has two tuning parameters: α, which determines the sparsity level of the network, and β, which controls the strength of resemblance between two neighboring precision estimators. In simulations and the empirical application we use the following procedure for tuning α and β: first, we set a grid of values $(\alpha,\beta) \in \{0, 0.25, 0.5, 1, 10, 30\}$. Second, we use the first 2/3 of the training data to estimate forecast combination weights and jointly tune α and β in the remaining 1/3 to yield the smallest value of the objective function, which is chosen to be either $\|\|\cdot\|\|_2$-loss of precision matrix for simulations in Subsection 6.1, or MSFE for simulations in Subsection 6.2 and the empirical application. Note that when $\beta = 0$, the optimization in (5.1) reduces to estimating

$\boldsymbol{\Theta}_{\varepsilon,i}$ using Algorithm 2 in each regime separately. Naturally, this incorporates the case when the structural break is strong and only the post-break data is used for producing forecast combination weights. When β is large, there are weak structural breaks in $\boldsymbol{\Theta}_{\varepsilon,i}$, and $\boldsymbol{\Theta}_{\varepsilon,i}$'s are estimated by using the data across different regimes. Section 7 provides more discussion on this in the context of our empirical application.

The smoothing function $\psi(\cdot)$ in (5.1) can be LASSO ($\psi = \sum_{l,q}|\cdot|$), Group LASSO ($\psi = \sum_q \|\cdot_q\|_2$), Ridge ($\psi = \sum_{l,q}(\cdot_{lq})^2$), Max norm penalty ($\psi = \sum_q \max_l|\cdot_{lq}|$). LASSO penalty encourages small changes in the precision matrix over time: when the lq-th element changes at two consecutive times, the penalty forces the rest of the elements of the precision to remain the same. Group LASSO penalty allows the entire graph to restructure at some time points. This penalty is useful for anomaly detection, since it can identify structural changes in the network structure. Ridge penalty allows the network to change smoothly over time. This penalty is less strict than the LASSO penalty: instead of encouraging the graphs to be exactly the same, it allows smooth transitions. For max-norm penalty, if an element of the precision matrix changes by ϵ in regime i, then other elements are allowed to change by up to ϵ without any additional penalty. This penalty allows to capture changes in the clusters of forecast errors, whereas keeping the rest of the network unchanged. In our empirical application we use Ridge penalty to accommodate smooth transitions of precision over time.

To estimate (5.1) we use the alternating direction method of multipliers (ADMM) algorithm described in details in Supplementary Appendix B. Once $\boldsymbol{\Theta}_{\varepsilon,i}$ is estimated, we combine estimated factors, loadings and precision matrix of the idiosyncratic components using Sherman-Morrison-Woodbury formula to estimate the final precision matrix of forecast errors and use it to compute optimal forecast combination weights. We call the aforementioned procedure Regime Dependent Factor GLASSO (RD-Factor GLASSO) and summarize

it in Algorithm 3.

Algorithm 3 RD-Factor GLASSO

1: Estimate $\widehat{\mathbf{f}}_t$ and $\widehat{\mathbf{B}}_t$ in (3.1). Get $\widehat{\mathbf{\Sigma}}_f$, $\widehat{\mathbf{\Theta}}_f$ and $\widehat{\boldsymbol{\varepsilon}}_t = \widehat{\mathbf{B}}\widehat{\mathbf{f}}_t - \mathbf{e}_t$.

2: Solve (5.1) using ADMM to get $\widehat{\mathbf{\Theta}}_{\varepsilon,i}$.

3: Use $\widehat{\mathbf{\Theta}}_{\varepsilon,i}$, $\widehat{\mathbf{\Theta}}_f$ and $\widehat{\mathbf{B}}$ from Steps 1-2 to get $\widehat{\mathbf{\Theta}}_i$:

$$\widehat{\mathbf{\Theta}}_i = \widehat{\mathbf{\Theta}}_{\varepsilon,i} - \widehat{\mathbf{\Theta}}_{\varepsilon,i}\widehat{\mathbf{B}}[\widehat{\mathbf{\Theta}}_f + \widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_{\varepsilon,i}\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Theta}}_{\varepsilon,i}. \tag{5.2}$$

4: Use $\widehat{\mathbf{\Theta}}_i$ to get forecast combination weights $\widehat{\mathbf{w}}_i = \frac{\widehat{\mathbf{\Theta}}_i \boldsymbol{\iota}_p}{\boldsymbol{\iota}_p' \widehat{\mathbf{\Theta}}_i \boldsymbol{\iota}_p}$.

We develop a scalable implementation of (5.1) for the RD-Factor GLASSO in Algorithm 3 through ADMM, which is extensively discussed in Supplementary Appendix B.

6 Monte Carlo

We divide the simulation results into two subsections. In the first subsection we study the consistency of the Factor GLASSO and RD-Factor GLASSO for estimating precision matrix and the combination weights. In the second subsection we evaluate the out-of-sample forecasting performance of combined forecasts in terms of MSFE. We compare the performance of forecast combinations based on the factor models in Algorithms 2, 3 with equal-weighted (EW) forecast combination, and combinations that use GLASSO without factor structure (Algorithm 1). Similarly to the literature on graphical models, all exercises use 100 Monte Carlo simulations.

6.1 Consistent Estimation of Forecast Combination Weights

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

The forecast errors are assumed to have the following structure:

$$\underbrace{\mathbf{e}_t}_{p\times 1} = \mathbf{B}\underbrace{\mathbf{f}_t}_{q\times 1} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T \tag{6.1}$$

$$\mathbf{f}_t = \phi_f \mathbf{f}_{t-1} + \boldsymbol{\zeta}_t, \tag{6.2}$$

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 (6.1) 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, $\mathbf{B} = (b)_{ij}$, where $(b)_{ij} = \rho^{|i-j|}$, $i, j \in \{1, \ldots, p\}$. We set $\rho = 0.2$, $\phi_f = 0.2$ and $\sigma_\zeta^2 = 1$. The specification in (6.1) leads to the low-rank plus sparse decomposition of the covariance matrix:

$$\mathbb{E}[\mathbf{e}_t \mathbf{e}_t'] = \mathbf{\Sigma} = \mathbf{B}\mathbf{\Sigma}_f \mathbf{B}' + \mathbf{\Sigma}_\varepsilon. \tag{6.3}$$

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 $\mathbf{A}_\varepsilon$ which represents the structure of the graph:

$$a_{\varepsilon,ij} = \begin{cases} 1, & \text{for } i \neq j \text{ with probability } \pi, \\ 0, & \text{otherwise,} \end{cases} \tag{6.4}$$

where $a_{\varepsilon,ij}$ denotes the i,j-th element of the adjacency matrix $\mathbf{A}_\varepsilon$. We set $a_{\varepsilon,ij} = a_{\varepsilon,ji} = 1$, 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 $\pi = 500/(pT^{0.8})$, which makes $s_T = \mathcal{O}(T^{0.05})$. The adjacency matrix has all diagonal elements equal to zero. To generate a sparse symmetric positive-definite precision matrix we use Scikit-Learn datasets package in Python (Pedregosa et al. (2011)). To control the magnitude of partial correlations, the value of the smallest coefficient is set to 0.1 and the value of the largest coefficient is set to 0.3.

Figure 3 shows 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 combination is obtained using the fact that diagonal covariance and precision matrices imply equal weights. 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 (2004)). As evidenced by Figure 3, Factor GLASSO demonstrates superior performance over EW and non-factor based model (GLASSO). 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). Also, note that the precision matrix estimated using the EW method also shows good convergence properties. However, in terms of estimating the combination weight, the performance of EW does not exhibit good convergence properties.

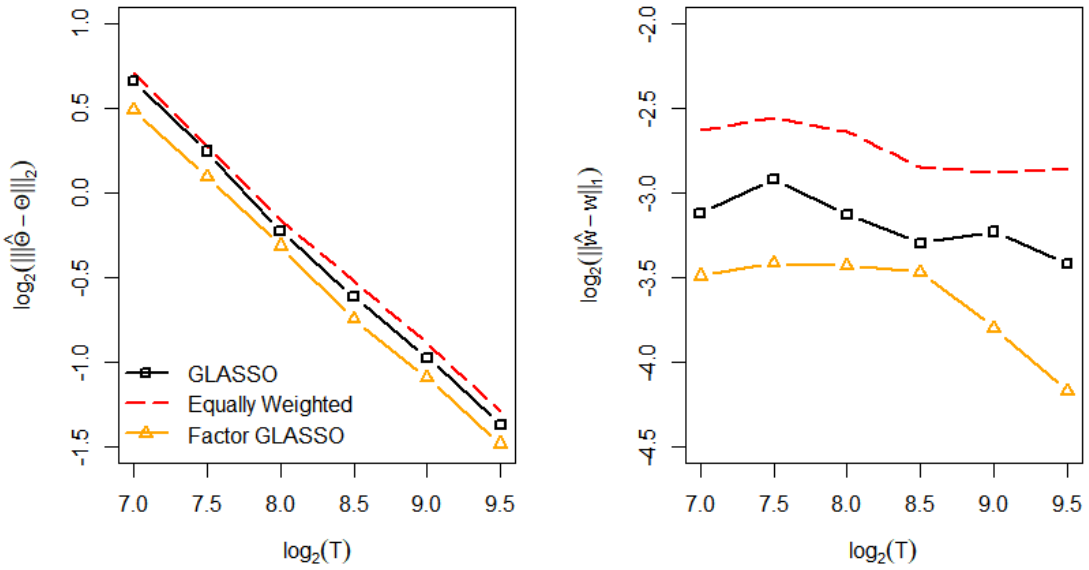

Figure 3: **Averaged errors of the estimators of Θ (left) and w on logarithmic scale (base 2).** $p = T^{0.85}$, $q = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.

To incorporate structural break in $\mathbf{\Theta}_\varepsilon$ we add the following modification to the DGP described above. We fix a single break point in the middle of the sample size, $T/2$: in the precision matrix of the idiosyncratic errors before the break, referred to as $\mathbf{\Theta}_{\varepsilon,1}$, the value of the largest coefficient is set to 0.4; whereas in the precision matrix of the idiosyncratic errors

after the break, $\mathbf{\Theta}_{\varepsilon,2}$, the value of the largest coefficient is set to 0.6. As a consequence, even though both matrices are still sparse, $\mathbf{\Theta}_{\varepsilon,2}$ has larger partial correlations. We use $\mathbf{\Theta}_{\varepsilon,1}$ and $\mathbf{\Theta}_{\varepsilon,2}$ to generate $\boldsymbol{\varepsilon}_t$ in (6.1). Figure 4 shows the performance of all models including RD-Factor GLASSO: accounting for the break in $\mathbf{\Theta}_{\varepsilon}$ significantly reduces the estimation error of precision matrix and combination weights. Notice that Factor GLASSO still outperforms GLASSO, although the dominance is less pronounced compared to the setup without structural breaks.

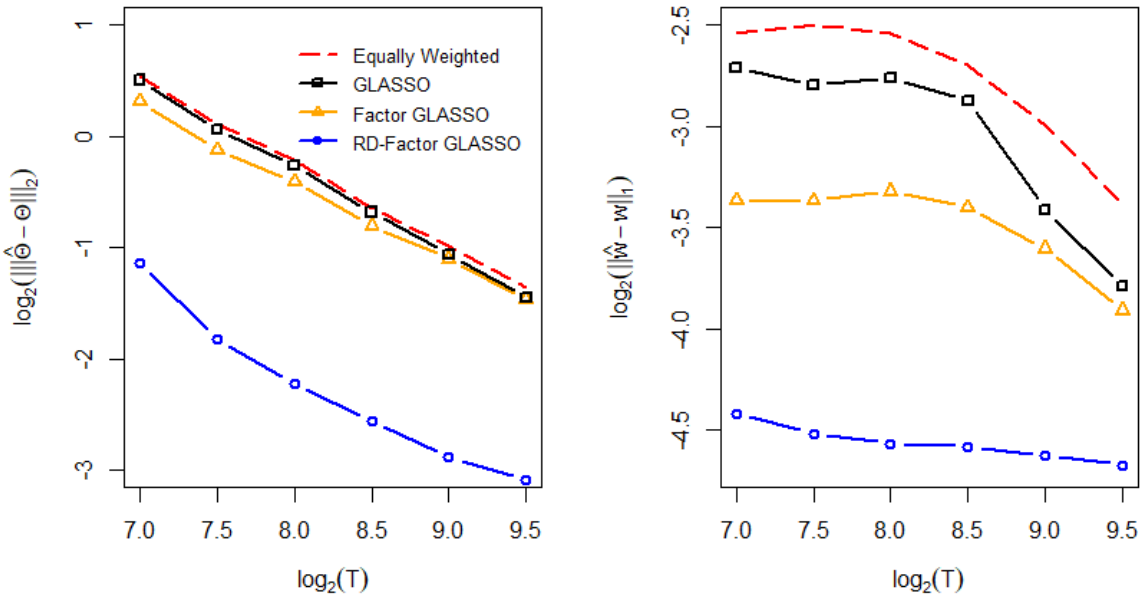

Figure 4: **Averaged errors of the estimators of Θ (left) and w on logarithmic scale (base 2).** $p = T^{0.85}$, $q = 2(\log(T))^{0.5}$, $s_T = \mathcal{O}(T^{0.05})$.

6.2 Comparing Performance of Forecast Combinations

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. Suppose the

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

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{g}_t + \mathbf{v}_t, \tag{6.5}$$

$$\mathbf{g}_t = \phi\mathbf{g}_{t-1} + \boldsymbol{\xi}_t, \tag{6.6}$$

$$y_{t+1} = \mathbf{g}_t'\boldsymbol{\alpha} + \sum_{s=1}^{\infty}\theta_s\epsilon_{t+1-s} + \epsilon_{t+1}, \tag{6.7}$$

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, $\mathbf{\Lambda}$ is an $N \times r$ matrix of factor loadings, $\mathbf{v}_t$ is an $N \times 1$ random vector following $\mathcal{N}(0, \sigma_v^2\mathbf{I}_N)$, 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

$$\theta_s = (1+s)^{c_1}c_2^s, \tag{6.8}$$

as in Hansen (2008). We set $c_1 \in \{0, 0.75\}$ and $c_2 = 0.9$. We set $N = 100$ and generate $r = 5$ factors. To create $\mathbf{\Lambda}$ in (6.5) we take the first r rows of an upper triangular matrix from a Cholesky decomposition of the $N \times N$ Toeplitz matrix parameterized by $\rho = 0.9$. The ranking of competing models was not very sensitive to varying values of ϕ, ρ, c_2, and r – the results examining sensitivity to a grid of 10 different AR(1) coefficients ϕ equidistant between 0 and 0.9, a grid of 10 different values of ρ equidistant between 0 and 0.9, $c_2 \in \{0.6, 0.7, 0.8, 0.9\}$, and $r = [1, \cdots, 7]$ are available upon request.

One-step ahead forecasts are estimated from the factor-augmented autoregressive (FAR)

models of orders k, l, denoted as FAR(k, l):

$$\hat{y}_{t+1} = \hat{\mu} + \hat{\kappa}_1 \hat{g}_{1,t} + \cdots + \hat{\kappa}_k \hat{g}_{k,t} + \hat{\psi}_1 y_t + \cdots + \hat{\psi}_l y_{t+1-l}, \tag{6.9}$$

where the factors $(\hat{g}_{1,t}, \ldots, \hat{g}_{k,t})$ are estimated from equation (6.5). 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. We set $K = 2$ and $L = 7$.

The total number of observations is T. The period for training the models is set to be $m_1 = T/2$ – this is used to train competing FAR models in (6.9). The remaining part of the sample, $m_2 = T/2$ is split as follows: the estimation window for training competing models (that is, EW, GLASSO, Factor GLASSO, and RD-Factor GLASSO) is set to be window $= m_2/2$. We roll the estimation window over the the test sample of the size $m_2/2$ to update all the estimates in each point of time $t = 1, \ldots, m/2$. Recall that q denotes the number of factors in the forecast errors as in equation (3.1).

Figure 5 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 5, the models that use the factor structure outperform EW combination and non-factor based counterparts for both values of c_1.

To incorporate structural break we add the following modification to the DGP described above. The period for training the models is set to be $m_1 = T/3$ – this is used to train competing FAR models in (6.9). The remaining part of the sample, $m_2 = 2 \times T/3$ is split as follows: the estimation window for training competing models is set to be window $= m_2/2$.

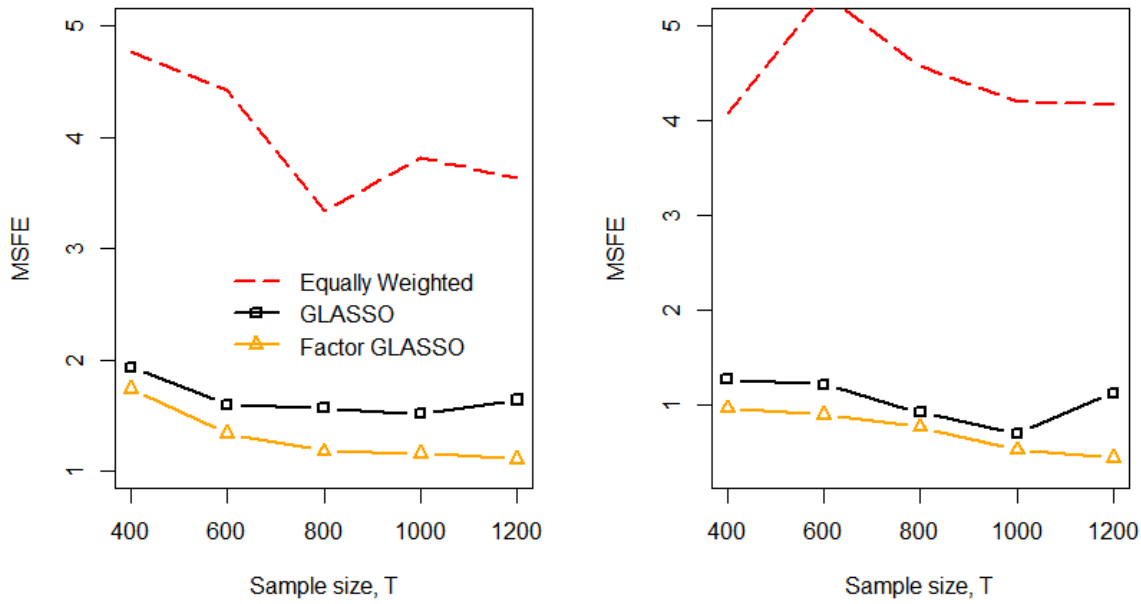

Figure 5: **Plots of the MSFE over the sample size T.** $c_1 = 0$ (left), $c_1 = 0.75$ (right), $c_2 = 0.9$, $N = 100$, $r = 5, \sigma_\xi = 1$, $L = 7$, $K = 2$, $p = 24$, $q = 3$, $\rho = 0.9$, $\phi = 0.8$.

We roll the estimation window over the the test sample of the size $m_2/2$. The break point is fixed at $1/2$ of the first estimation window. Before the break, when generating θ_s in (6.8) we set $c_2 = 0.3$, and after the break $c_2 = 0.9$. All other parameters stay unchanged. Figure 6 shows the performance of all models including RD-Factor GLASSO: similarly to the conclusions in the previous subsection, accounting for the break significantly reduces MSFE of the combined forecast.

7 Application to Combining ECB SPF Forecasts

We use quarterly forecasts on the expected rates of inflation, real GDP growth and unemployment rate in the Euro area published by the ECB. The raw data records 119

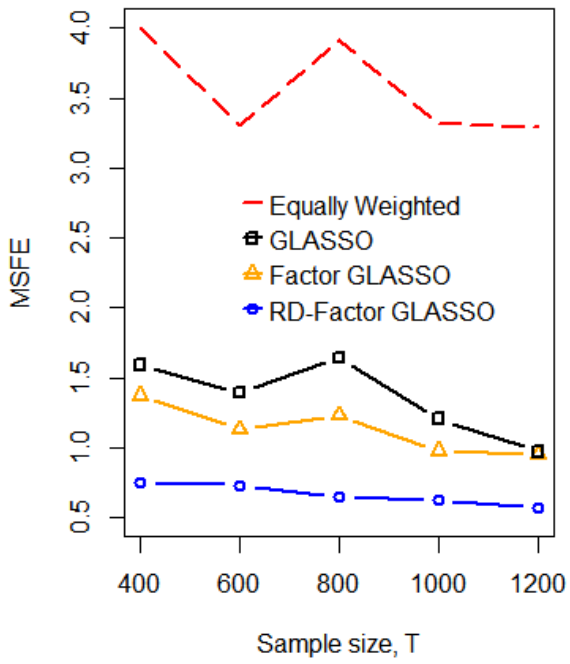

Figure 6: **Plots of the MSFE over the sample size T.** $c_1 = 0.75$, $c_2 = 0.3$ (before the break), $c_2 = 0.9$ (after the break), $N = 100$, $r = 5, \sigma_\xi = 1$, $L = 7$, $K = 2$, $p = 24$, $q = 3$, $\rho = 0.9$, $\phi = 0.8$.

forecasters in total, but the panel is highly unbalanced with many missing values due to entry and exit in the long span. We follow Shi et al. (2020) to obtain most qualified forecasters: first, we filter out irregular respondents if they missed more than 50% of the observations; second, we use a random forest imputation algorithm (Stekhoven (2022); Stekhoven and Buhlmann (2012)) to interpolate the remaining missing values. We consider the forecasts of three main economic indicators: (1) Real GDP growth defined as the year-on-year (YoY) percentage change of real GDP, based on standardized European System of National and Regional Accounts (ESA) definition. The time period under consideration is 1999:Q3-2019Q4 (which yields the sample size $T = 82$), the final number of forecasters is $p = 30$, and the prediction horizon is 2-quarters ahead. (2) Inflation which is defined as the YoY percentage

change of the Harmonised Index of Consumer Prices (HICP) published by Eurostat. The time period under consideration is 1999:Q4-2019Q4 (which yields the sample size $T = 81$), the final number of forecasters is $p = 50$, and the prediction horizon is 2-quarters ahead. (3) Unemployment rate which refers to Eurostat's definition and it is calculated as percentage of the labor force. The time period under consideration is 1999:Q3-2019Q4 (which yields the sample size $T = 82$), the final number of forecasters is $p = 43$, and the prediction horizon is 3-quarters ahead.

We consider three choices of the training sample: $m \in \{20, 30, 40, 50\}$, the estimation window is rolled over the test sample to update the estimates in each point of time. 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). In the majority of the cases the optimal number of factors was estimated to be equal to 1. To explore the benefits of using Factor GLASSO and RD-Factor GLASSO for forecast error quantification, we consider several alternative estimators of covariance/precision matrix of the idiosyncratic component in (4.1): (1) linear shrinkage estimator of covariance developed by Ledoit and Wolf (2004) further referred to as Factor LW; (2) nonlinear shrinkage estimator of covariance by Ledoit and Wolf (2017) (Factor NLW); (3) POET (Fan et al. (2013)); (4) constrained ℓ_1-minimization for inverse matrix estimator, CLIME (Cai et al. (2011)) (Factor CLIME); (5) nodewise regression developed by Meinshausen and Bühlmann (2006) (Factor MB). To examine the benefits of imposing sparsity on $\mathbf{\Theta}_\varepsilon$ we also include the factor model without sparsity assumption on the idiosyncratic error precision matrix (referred to as Not Sparse) – this corresponds to imposing $\tau = 0$ in (4.2). Our benchmark is the simple average with equal weights on all forecasters (referred to as EW). Going back to the discussion in Section 5 regarding setting $\beta = 0$ in equation (5.1): as we pointed out,

this corresponds to using only post-break sample for estimation which is suboptimal since the value of β is already chosen optimally from the grid that includes $\beta = 0$ to minimize the MSFE. Hence, by construction, RD-Factor GLASSO is superior to using only post-break data.

For RD-Factor GLASSO we consider the following structural break points associated with a global financial crisis of 2007-09: for real GDP series the break is taken to be 2009:Q1, for inflation series the break is 2008:Q2, and for unemployment rate – 2008:Q2. We chose these points based on the behavior of the actual series which exhibited a change in volatility following the break. This corresponds to one structural break for each series, or, in terms of the notations used in (5.1), $N = 2$ (two regimes).

Table 1 compares the performance of Factor GLASSO and RD-Factor GLASSO with the competitors for predicting three macroeconomic indicators for Euro-area using a combination of ECB SPF forecasts. For the inflation series training the models is computationally challenging due to $p \geq T$, hence the numerical routine used in several models (Factor NLW and Factor CLIME) failed to converge and we were not able to get the final estimates despite making appropriate changes to the tuning parameters used in NLW and CLIME. Furthermore, GLASSO also had difficulty in handling inflation series since for all m almost all off-diagonal elements of $\mathbf{\Theta}$ were shrunk to zero. As a consequence, for the inflation series GLASSO produced the same results as EW.

There are three main findings that we learn from analyzing Table 1: **(1)** for all series factor-based models outperform non-factor ones (such as EW and GLASSO). This means that incorporating the factor structure in the forecast errors improves forecasting performance. **(2)** for all series the Not Sparse model provides the worst performance. This means that the factor structure per se is not sufficient to achieve performance gains over EW, hence,

	GLASSO	Factor GLASSO	Not Sparse	Factor LW	Factor NLW	POET	Factor CLIME	Factor MB	RD Factor GLASSO
Real GDP Growth									
$m = 50$	1.18E-04	7.61E-05	3.67E-04	1.13E-04	1.59E-04	1.13E-04	1.09E-04	7.58E-05	6.45E-05
Ratio	1.61	1.04	5.00	1.54	2.17	1.54	1.49	1.03	**0.88**
p-val	0.94	0.65	1.00	0.93	0.99	0.89	0.90	0.52	0.08
$m = 40$	4.73E-04	9.58E-05	1.37E-03	4.17E-04	5.37E-04	1.40E-04	5.95E-04	1.25E-04	9.49E-05
Ratio	4.89	0.99	14.14	4.32	5.56	1.45	6.16	1.29	**0.98**
p-val	0.95	0.48	0.98	0.93	0.95	0.95	0.95	0.86	0.45
$m = 30$	3.54E-04	2.19E-04	4.61E-03	3.01E-04	4.39E-04	2.90E-04	4.49E-04	2.83E-03	1.76E-04
Ratio	1.37	0.85	6.29	1.16	1.69	1.12	1.73	10.90	**0.68**
p-val	0.78	0.04	1.00	0.63	0.79	0.58	0.80	0.95	0.00
$m = 20$	4.08E-04	1.96E-04	8.47E-04	3.19E-04	3.79E-04	2.41E-04	4.50E-04	2.02E-04	1.89E-04
Ratio	1.75	0.84	3.64	1.37	1.63	1.03	1.93	0.87	**0.81**
p-val	0.86	0.03	0.97	0.74	0.91	1.00	0.84	0.29	0.04
Inflation									
$m = 50$	3.06E-04	2.18E-04	2.48E-03	2.19E-04	NaN	3.87E-04	NaN	2.69E-04	1.61E-04
Ratio	1.00	0.71	8.10	0.72		1.26		0.88	**0.53**
p-val	1.00	9.96E-03	0.97	5.17E-03		0.94		0.06	2.54E-04
$m = 40$	3.75E-04	3.54E-04	4.58E-04	3.03E-04	NaN	3.75E-04	NaN	4.33E-04	2.37E-04
Ratio	1.00	0.95	1.22	0.81		1.00		1.15	**0.63**
p-val	0.96	0.40	0.74	0.07		0.98		0.89	4.37E-03
$m = 30$	5.17E-04	5.10E-04	5.47E-04	4.53E-04	NaN	5.17E-04	NaN	5.25E-04	2.80E-04
Ratio	1.00	0.99	1.06	0.88		1.00		1.02	**0.54**
p-val	0.76	0.46	0.65	0.11		0.90		0.92	0.01
$m = 20$	5.37E-04	4.96E-04	5.43E-04	4.75E-04	NaN	5.61E-04	NaN	5.36E-04	2.85E-04
Ratio	1.00	0.92	1.01	0.88		1.05		1.00	**0.53**
p-val	0.87	0.13	0.56	1.75E-02		0.67		9.75E-04	3.37E-03
Unemployment Rate									
$m = 50$	8.20E-03	5.59E-03	4.04E-02	5.40E-03	4.49E-03	8.56E-03	5.04E-03	7.01E-03	5.54E-03
Ratio	0.92	0.63	4.54	0.61	**0.50**	0.96	0.57	0.79	0.62
p-val	9.04E-06	2.53E-13	0.91	2.22E-15	1.15E-16	0.01	2.16E-12	9.93E-09	0.29
$m = 40$	7.76E-03	5.17E-03	1.89E-02	5.22E-03	4.66E-03	8.23E-03	4.35E-03	7.12E-03	6.69E-03
Ratio	0.93	0.62	2.28	0.63	0.56	0.99	**0.52**	0.86	0.77
p-val	1.21E-06	7.22E-15	0.89	8.00E-14	2.43E-14	0.14	3.11E-15	1.35E-08	0.41
$m = 30$	7.04E-03	4.26E-03	1.44E-02	4.41E-03	4.51E-03	7.22E-03	3.46E-03	6.61E-03	6.34E-03
Ratio	0.96	0.58	1.97	0.60	0.62	0.99	**0.47**	0.90	0.77
p-val	2.54E-03	6.29E-14	0.84	7.99E-15	1.18E-14	0.06	1.11E-15	1.75E-08	0.44
$m = 20$	6.85E-03	4.93E-03	1.58E-02	4.86E-03	5.41E-03	7.03E-03	3.79E-03	6.45E-03	6.15E-03
Ratio	1.00	0.72	2.30	0.71	0.79	1.03	**0.55**	0.94	0.82
p-val	0.49	9.47E-10	0.85	1.83E-10	1.15E-08	0.73	2.00E-16	7.87E-06	0.42

Table 1: **Prediction of Quarterly Macroeconomic Variables for Euro-area Using ECB SPF Forecasts.** MSFEs of competing methods are reported for each value of m, where m indicates the length of the training window. Ratio indicates the ratio to MSFE of the Equal-Weighted combined forecast. The best-performing models are in bold. P-values are computed according to the Diebold–Mariano test. Small p-values indicate the ratio is significantly smaller than one. Factor GLASSO and RD-Factor GLASSO are our proposed models.

it is necessary to impose sparsity on the precision matrix of the idiosyncratic components. **(3)** RD-Factor GLASSO is the best-performing model for real GDP and inflation series, and Factor GLASSO is the second-best performing model. For the unemployment rate, Factor GLASSO outperforms RD-Factor GLASSO. This result is supported by the behavior observed in the actual series: real GDP and inflation exhibit a clear break in variance following the global financial crisis of 2007-09, however this is not the case for the unemployment rate series that did not have noticeable changes in variance throughout the whole sample period.

8 Conclusions

In this paper we develop a unified framework to generalize network inference under a factor structure in the presence of structural breaks. We overcome the challenge of using graphical models under the factor structure and provide a simple approach that allows practitioners to combine a large number of forecasts when experts tend to make common mistakes. Using pre- and post-break data, our new approach to forecast combinations breaks down forecast errors into common and unique parts which improves the accuracy of the combined forecast. Our first algorithm, Factor GLASSO, is shown to consistently estimate forecast combination weights and MSFE. For the ease of practical use we develop a scalable optimization procedure for RD-Factor GLASSO, based on the ADMM. The empirical application to forecasting macroeconomic series using the data of the ECB Survey of Professional Forecasters shows that incorporating (i) factor structure in the forecast errors together with (ii) sparsity in the precision matrix of the idiosyncratic components and (iii) regime-dependent combination weights improves the performance of a combined forecast.

References

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221. 8, 30

Bates, J. M. and Granger, C. W. J. (1969). The combination of forecasts. *Operations Research*, 20(4):451–468. 1, 10

Brownlees, C., Nualart, E., and Sun, Y. (2018). Realized networks. *Journal of Applied Econometrics*, 33(7):986–1006. 12

Cai, T., Liu, W., and Luo, X. (2011). A constrained l1-minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607. 30

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. 37

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304. 9

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. 7

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583. 1

Diebold, F. 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. 1, 8

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4):603–680. 15, 30

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383–1414. 16

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441. 2, 5, 6, 7

Hallac, D., Park, Y., Boyd, S., and Leskovec, J. (2017). Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 205–213, New York, NY, USA. ACM. 18

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350. 26

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. 7, 17

Koike, Y. (2020). De-biased Graphical LASSO for high-frequency data. *Entropy*, 22(4):456. 12

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411. 24, 30

Ledoit, O. and Wolf, M. (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388. 30

Lee, T.-H. and Seregina, E. (2020). Optimal portfolio using Factor Graphical LASSO. *arXiv:2011.00435*. 17, 38

Lu, J., Kolar, M., and Liu, H. (2015). Post-regularization inference for time-varying nonparanormal graphical models. *Journal of Machine Learning Research*, 18:203:1–203:78. 18

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462. 2, 30

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239. 41

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 23

Ravikumar, P., J. Wainwright, M., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980. 5

Shi, Z., Su, L., and Xie, T. (2020). High dimensional forecast combinations under latent structures. *arXiv:2010.09477*. 29

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. 11

Stekhoven, D. J. (2022). *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.5. 29

Stekhoven, D. J. and Buhlmann, P. (2012). Missforest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118. 29

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. 8, 10, 25

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, 80(2):295–319. 18

Supplemental Appendix

Appendix A Proof of Theorem 1

We first present a lemma which is used in the proof.

Lemma 1.

(a) $|||\mathbf{\Theta}|||_1 = \mathcal{O}(d_T)$.

(b) $a \geq c > 0$, *where* a *was defined in Section 3 and* c *was defined in Assumption* ***(A.1)*** *(ii).*

(c) $|\widehat{a} - a| = \mathcal{O}_P(\varrho_{1T} d_T s_T)$, *where* $\widehat{a}$ *was defined in Section 3.*

Proof.

(a) To prove part (a) we use the following matrix inequality which holds for any $\mathbf{A} \in \mathcal{S}_p$:

$$|||\mathbf{A}|||_1 = |||\mathbf{A}|||_\infty \leq \sqrt{d(\mathbf{A})}|||\mathbf{A}|||_2, \tag{A.1}$$

where $d(\mathbf{A})$ was defined in Section 4. The proof of (A.1) is a straightforward consequence of the Schwarz inequality.

Sherman-Morrison-Woodbury formula together with (A.1) and Assumptions **(B.1)**-**(B.3)** yield:

$$\begin{aligned} |||\mathbf{\Theta}|||_1 &\leq |||\mathbf{\Theta}_\varepsilon|||_1 + \left|\left|\left|\mathbf{\Theta}_\varepsilon \mathbf{B}[\mathbf{\Theta}_f + \mathbf{B}'\mathbf{\Theta}_\varepsilon \mathbf{B}]^{-1}\mathbf{B}'\mathbf{\Theta}_\varepsilon\right|\right|\right|_1 \\ &= \mathcal{O}(\sqrt{d_T}) + \mathcal{O}\left(\sqrt{d_T} \cdot p \cdot \frac{1}{p} \cdot \sqrt{d_T}\right) = \mathcal{O}(d_T). \end{aligned} \tag{A.2}$$

(b) Under Assumption **(A.1)**:

$$a = \boldsymbol{\iota}_p' \mathbf{\Theta} \boldsymbol{\iota}_p / p \geq c > 0.$$

(c) Using the Hölders inequality, we have

$$|\widehat{a} - a| = \left|\frac{\boldsymbol{\iota}_p'(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\boldsymbol{\iota}_p}{p}\right| \leq \frac{\left\|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})\boldsymbol{\iota}_p\right\|_1 \|\boldsymbol{\iota}_p\|_\infty}{p} \leq \left|\left|\left|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\right|\right|\right|_1$$

$$= \mathcal{O}_P(\varrho_{1T} d_T s_T) = o_P(1),$$

where the last rate is obtained using the assumptions of Theorem 1.

□

A.1 Proof of Theorem 1

First, note that the forecast combination weight can be written as

$$\begin{aligned}\widehat{\mathbf{w}} - \mathbf{w} &= \frac{\Big((a\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p) - (\hat{a}\boldsymbol{\Theta}\boldsymbol{\iota}_p)\Big)/p}{\hat{a}a} \\ &= \frac{\Big((a\widehat{\boldsymbol{\Theta}}\boldsymbol{\iota}_p) - (a\boldsymbol{\Theta}\boldsymbol{\iota}_p) + (a\boldsymbol{\Theta}\boldsymbol{\iota}_p) - (\hat{a}\boldsymbol{\Theta}\boldsymbol{\iota}_p)\Big)/p}{\hat{a}a}.\end{aligned}$$

As shown in Callot et al. (2019), the above can be rewritten as

$$\|\widehat{\mathbf{w}} - \mathbf{w}\|_1 \leq \frac{a\frac{\left\|(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1}{p} + |a - \widehat{a}|\frac{\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1}{p}}{|\widehat{a}|a}. \tag{A.3}$$

Prior to bounding the terms in (A.3), we first present an inequality which is used in the derivations. Let $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{v} \in \mathbb{R}^{p\times 1}$. Also, let $\mathbf{A}_j$ and $\mathbf{A}_j'$ be a $p \times 1$ and $1 \times p$ row and column vectors in $\mathbf{A}$, respectively.

$$\begin{aligned}\|\mathbf{A}\mathbf{v}\|_1 &= |\mathbf{A}_1'\mathbf{v}| + \ldots + \left|\mathbf{A}_p'\mathbf{v}\right| \leq \|\mathbf{A}_1\|_1\|\mathbf{v}\|_\infty + \ldots + \|\mathbf{A}_p\|_1\|\mathbf{v}\|_\infty \\ &= \left(\sum_{j=1}^{p}\|\mathbf{A}_j\|_1\right)\|\mathbf{v}\|_\infty \leq p\max_j|\mathbf{A}_j|_1\|\mathbf{v}\|_\infty.\end{aligned} \tag{A.4}$$

Hölders inequality was used to obtain each inequality in (A.4). If $\mathbf{A} \in \mathcal{S}_p$, then the last expression can be further reduced to $p\|\|\mathbf{A}\|\|_1\|\mathbf{v}\|_\infty$.

Let us now bound the right-hand side of (A.3). In the numerator we have:

$$\frac{\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1}{p} \leq \|\|\boldsymbol{\Theta}\|\|_1 = \mathcal{O}_P(\varrho_{1T}d_T s_T), \tag{A.5}$$

the rates was derived in Lee and Seregina (2020), and the inequality follows from (A.4).

$$\frac{\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1}{p} \leq |||\boldsymbol{\Theta}|||_1 = \mathcal{O}(d_T), \tag{A.6}$$

where the rate follows from Lemma 1 (a) and the inequality is obtained from (A.4). Combining (A.5), (A.6), and Lemma 1 (c) we get:

$$\begin{aligned} a\frac{\left\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{\iota}_p\right\|_1}{p} + |a - \widehat{a}|\frac{\|\boldsymbol{\Theta}\boldsymbol{\iota}_p\|_1}{p} &= \mathcal{O}(1)\cdot\mathcal{O}_P(\varrho_{1T}d_T s_T) + \mathcal{O}_P(\varrho_{1T}d_T s_T)\cdot\mathcal{O}(d_T) \\ &= \mathcal{O}_P(\varrho_{1T}d_T^2 s_T) = o_P(1), \end{aligned} \tag{A.7}$$

where the last equality holds under the assumptions of Theorem 1.

For the denominator of (A.3) it easy to see that $|\widehat{a}|a = \mathcal{O}_P(1)$ using the results of Lemma 1 (b).

For the MSFE part of Theorem 1, using Lemma 1 (b)-(c), we get

$$\left|\frac{\hat{a}^{-1}}{a^{-1}} - 1\right| = \frac{|a - \hat{a}|}{|\hat{a}|} = \mathcal{O}_P(\varrho_{1T}d_T s_T) = o_P(1),$$

where the last rate is obtained using the assumptions of Theorem 1.

Appendix B Implementation via ADMM Algorithm

To enable practical implementation of the RD-Factor GLASSO, we develop an optimization procedure using ADMM algorithm to solve the convex optimization problem in (5.1).

First, we need to reformulate the unconstrained problem in (5.1) as a constrained problem which can be solved using ADMM:

$$\{\widehat{\boldsymbol{\Theta}}_{\varepsilon,i}\}_{i=1}^{N} = \arg\min_{\{\boldsymbol{\Theta}_{\varepsilon,i}\}_{i=1}^{N}} \sum_{i=1}^{N} n_i \Big[\operatorname{trace}\Big(\widehat{\boldsymbol{\Sigma}}_{\varepsilon,i}\boldsymbol{\Theta}_{\varepsilon,i}\Big) - \log\det\boldsymbol{\Theta}_{\varepsilon,i}\Big] + \alpha\|\boldsymbol{\Theta}_{\varepsilon,i}\|_{\text{od},1} \tag{B.1}$$
$$+ \beta\sum_{i=2}^{N}\psi(\boldsymbol{\Theta}_{\varepsilon,i} - \boldsymbol{\Theta}_{\varepsilon,i-1})$$

$$\text{s.t. } \mathbf{Z}_{i,0} = \boldsymbol{\Theta}_{\varepsilon,i}, \text{ for } i = 1,\ldots,N \tag{B.2}$$

$$\Big(\mathbf{Z}_{i-1,1}, \mathbf{Z}_{i,2}\Big) = \Big(\boldsymbol{\Theta}_{\varepsilon,i-1}, \boldsymbol{\Theta}_{\varepsilon,i}\Big), \text{ for } i = 2,\ldots,N. \tag{B.3}$$

Let $\mathbf{Z} = \Big\{\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2\Big\} = \Big\{\Big(\mathbf{Z}_{1,0},\ldots,\mathbf{Z}_{N,0}\Big), \Big(\mathbf{Z}_{1,1},\ldots,\mathbf{Z}_{N-1,1}\Big), \Big(\mathbf{Z}_{2,2},\ldots,\mathbf{Z}_{N,2}\Big)\Big\}$.

Let $\mathbf{U} = \Big\{\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2\Big\} = \Big\{\Big(\mathbf{U}_{1,0},\ldots,\mathbf{U}_{N,0}\Big), \Big(\mathbf{U}_{1,1},\ldots,\mathbf{U}_{N-1,1}\Big), \Big(\mathbf{U}_{2,2},\ldots,\mathbf{U}_{N,2}\Big)\Big\}$ be the scaled dual variable and $\rho > 0$ is the ADMM penalty parameter. Now we can use scaled ADMM to write down the augmented Lagrangian:

$$\mathcal{L}_\rho(\boldsymbol{\Theta}_\varepsilon, \mathbf{Z}, \mathbf{U}) = \sum_{i=1}^{T} n_i \Big[\operatorname{trace}\Big(\widehat{\boldsymbol{\Sigma}}_{\varepsilon,i}\boldsymbol{\Theta}_{\varepsilon,i}\Big) - \log\det\boldsymbol{\Theta}_{\varepsilon,i}\Big] + \alpha\|\mathbf{Z}_{i,0}\|_{od,1} \tag{B.4}$$
$$+ \beta\sum_{i=2}^{T}\psi(\mathbf{Z}_{i,2} - \mathbf{Z}_{i-1,1})$$
$$+ \Big(\frac{\rho}{2}\Big)\sum_{i=1}^{T}\Big(\|\boldsymbol{\Theta}_{\varepsilon,i} - \mathbf{Z}_{i,0} + \mathbf{U}_{i,0}\|_F^2 - \|\mathbf{U}_{i,0}\|_F^2\Big)$$
$$+ \Big(\frac{\rho}{2}\Big)\sum_{i=2}^{T}\Big(\|\boldsymbol{\Theta}_{\varepsilon,i-1} - \mathbf{Z}_{i-1,1} + \mathbf{U}_{i-1,1}\|_F^2 - \frac{\rho}{2}\|\mathbf{U}_{i-1,1}\|_F^2$$
$$+ \|\boldsymbol{\Theta}_{\varepsilon,i} - \mathbf{Z}_{i,2} + \mathbf{U}_{i,2}\|_F^2 - \|\mathbf{U}_{i,2}\|_F^2\Big).$$

Let k denote the iteration number, then ADMM consists of the following iterative updates:

$$\boldsymbol{\Theta}_{\varepsilon,i}^{k+1} \equiv \arg\min_{\boldsymbol{\Theta}_\varepsilon \succ 0} \mathcal{L}_\rho(\boldsymbol{\Theta}_\varepsilon, \mathbf{Z}^k, \mathbf{U}^k), \tag{B.5}$$

$$\mathbf{Z}^{k+1} = \begin{bmatrix} \mathbf{Z}_0^{k+1} \\ \mathbf{Z}_1^{k+1} \\ \mathbf{Z}_2^{k+1} \end{bmatrix} \equiv \arg\min_{\mathbf{Z}_0, \mathbf{Z}_1, \mathbf{Z}_2} \mathcal{L}_\rho(\boldsymbol{\Theta}_\varepsilon^{k+1}, \mathbf{Z}, \mathbf{U}^k), \tag{B.6}$$

$$\mathbf{U}^{k+1} = \begin{bmatrix} \mathbf{U}_0^{k+1} \\ \mathbf{U}_1^{k+1} \\ \mathbf{U}_2^{k+1} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{U}_0^{k} \\ \mathbf{U}_1^{k} \\ \mathbf{U}_2^{k} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Theta}_\varepsilon^{k+1} - \mathbf{Z}_0^{k+1} \\ (\boldsymbol{\Theta}_{\varepsilon,1}^{k+1}, \ldots, \boldsymbol{\Theta}_{\varepsilon,N-1}^{k+1}) - \mathbf{Z}_1^{k+1} \\ (\boldsymbol{\Theta}_{\varepsilon,2}^{k+1}, \ldots, \boldsymbol{\Theta}_{\varepsilon,N}^{k+1}) - \mathbf{Z}_2^{k+1} \end{bmatrix}. \tag{B.7}$$

The Z step:
The updating rule in (B.6) is easily recognized to be the element-wise soft thresholding operator. However, we need to split it into two updates since $(\mathbf{Z}_1, \mathbf{Z}_2)$ have to be updated jointly. Therefore, the update for $\mathbf{Z}_{i,0}^{k+1}$ will be:

$$\mathbf{Z}_{i,0}^{k+1} \equiv S_{\alpha/\rho}(\boldsymbol{\Theta}_{\varepsilon,i}^{k+1} + \mathbf{U}_{i,0}^k), \tag{B.8}$$

where $S_{\alpha/\rho}(\cdot)$ is the element-wise soft-thresholding operator for $i \neq j$.

We will solve a separate update for each $(\mathbf{Z}_{i,2}, \mathbf{Z}_{i-1,1})$ pair for $i = 2, \ldots, N$:

$$\begin{aligned}(\mathbf{Z}_{i,2}^{k+1}, \mathbf{Z}_{i-1,1}^{k+1}) = \arg\min_{\mathbf{Z}_{i,2}, \mathbf{Z}_{i-1,1}} \left(\frac{\rho}{2}\right) \Big(& \|\boldsymbol{\Theta}_{\varepsilon,i} - \mathbf{Z}_{i,2} + \mathbf{U}_{i,2}\|_F^2 \\ & + \|\boldsymbol{\Theta}_{\varepsilon,i-1} - \mathbf{Z}_{i-1,1} + \mathbf{U}_{i-1,1}\|_F^2 + \beta\psi(\mathbf{Z}_{i,2} - \mathbf{Z}_{i-1,1}) \Big).\end{aligned} \tag{B.9}$$

Note that (B.9) is guaranteed to converge to a fixed point since it can be written as a proximal operator:

$$(\mathbf{Z}_{i,2}^{k+1}, \mathbf{Z}_{i-1,1}^{k+1}) = \text{prox}_{\frac{\beta}{\rho}\psi(\cdot)}\Big(\boldsymbol{\Theta}_{\varepsilon,i} + \mathbf{U}_{i,2}, \boldsymbol{\Theta}_{\varepsilon,i-1} + \mathbf{U}_{i-1,1}\Big) \tag{B.10}$$

Remark 1. *A proximal operator of the scaled function νf, where $\nu > 0$ can be expressed as:*

$$prox_{\nu f}(v) = \underset{x}{\operatorname{argmin}} \left(f(x) + \frac{1}{2\nu} \|x - v\|_2^2 \right),$$

where f is a closed proper convex function. Note:

$$prox_{\nu f}(v) \approx v - \tau \nabla f(v).$$

Parikh and Boyd (2014) show that the fixed points of the proximal operator of f are precisely the minimizers of f, i.e. $prox_{\nu f}(x^\star) = x^\star$ if and only if $x^\star$ minimizes f.

The $\boldsymbol{\Theta}$ step:
The updating rule in (B.5) can be further simplified to obtain a closed-form solution. Rewrite (B.5):

$$\boldsymbol{\Theta}_{\varepsilon}^{k+1} = \underset{\Theta_\varepsilon \succ 0}{\operatorname{argmin}} \operatorname{trace}\left(\widehat{\boldsymbol{\Sigma}}_i \boldsymbol{\Theta}_{\varepsilon,i} \right) - \log\det \boldsymbol{\Theta}_{\varepsilon,i} + \frac{1}{2\eta} \left\| \boldsymbol{\Theta}_{\varepsilon,i} - \mathbf{A}^k \right\|_F^2, \tag{B.11}$$

where $\mathbf{A}^k = \dfrac{\mathbf{Z}_{i,0}^k + \mathbf{Z}_{i-1,1}^k + \mathbf{Z}_{i,2}^k - \mathbf{U}_{i,0}^k - \mathbf{U}_{i-1,1}^k - \mathbf{U}_{i,2}^k}{3}$, and $\eta = \frac{n_i}{3\rho}$.

Take the gradient of the updating rule in (B.11) in order to get an analytical solution:

$$\widehat{\boldsymbol{\Sigma}}_{\varepsilon,i} - \boldsymbol{\Theta}_{\varepsilon,i}^{-1,(k+1)} + \frac{1}{\eta}\left(\boldsymbol{\Theta}_{\varepsilon,i}^{k+1} - \mathbf{A}^k \right) = 0, \tag{B.12}$$

$$\frac{1}{\eta}\boldsymbol{\Theta}_{\varepsilon,i}^{k+1} - \boldsymbol{\Theta}_{\varepsilon}^{-1,(k+1)} = \frac{1}{\eta}\mathbf{A}^k - \widehat{\boldsymbol{\Sigma}}_{\varepsilon,i}. \tag{B.13}$$

Equation (B.13) implies that $\boldsymbol{\Theta}_{\varepsilon,i}^{k+1}$ and $\frac{1}{\eta}\mathbf{A}^k - \widehat{\boldsymbol{\Sigma}}_{\varepsilon,i}$ share the same eigenvectors.

Let $\mathbf{Q}_i \boldsymbol{\Lambda}_i \mathbf{Q}_i'$ be the eigendecomposition of $\frac{1}{\eta}\mathbf{A}^k - \widehat{\boldsymbol{\Sigma}}_{\varepsilon,i}$, where $\boldsymbol{\Lambda}_i = \operatorname{diag}(\lambda_{1,i}, \ldots, \lambda_{p,i})$, and $\mathbf{Q}_i'\mathbf{Q}_i = \mathbf{Q}_i\mathbf{Q}_i' = \mathbf{I}$.[1] Pre-multiply (B.13) by $\mathbf{Q}_i'$ and post-multiply it by $\mathbf{Q}_i$:

$$\frac{1}{\eta}\widetilde{\boldsymbol{\Theta}}_{\varepsilon,i}^{k+1} - \widetilde{\boldsymbol{\Theta}}_{\varepsilon,i}^{-1,(k+1)} = \boldsymbol{\Lambda}_i. \tag{B.14}$$

[1]Note that in practice we need to check that $\mathbf{A}^k$ is symmetric. If it is not, then we can define $\widetilde{\mathbf{A}} \equiv \frac{\mathbf{A}^k + (\mathbf{A}^k)'}{2}$ and use it in the described algorithm instead of $\mathbf{A}^k$. Since $\boldsymbol{\Theta}_i$ is symmetric the results will not be affected.

Now construct a diagonal solution of (B.14):

$$\frac{1}{\eta}\tilde{v}_{j,i} - \frac{1}{\tilde{v}_{j,i}} = \lambda_{j,i}, \tag{B.15}$$

where $\tilde{v}_{i,j}$ denotes the j-th eigenvalue of $\widetilde{\boldsymbol{\Theta}}_{\varepsilon,i}$. Solving for $\tilde{v}_{j,i}$ we get:

$$\tilde{v}_{j,i} = \frac{\lambda_{j,i} + \sqrt{\lambda_{j,i}^2 + \frac{4}{\eta}}}{2\eta^{-1}}. \tag{B.16}$$

Now we can calculate $\boldsymbol{\Theta}_{\varepsilon,i}^{k+1}$ which satisfies the optimality condition in (B.14):

$$\boldsymbol{\Theta}_{\varepsilon,i}^{k+1} = \frac{1}{2\eta^{-1}}\mathbf{Q}_i\Big(\boldsymbol{\Lambda}_i + \sqrt{\boldsymbol{\Lambda}_i^2 + 4\eta^{-1}\mathbf{I}}\Big)\mathbf{Q}_i'. \tag{B.17}$$

Use the definition of $\eta = \frac{n_i}{3\rho}$:

$$\boldsymbol{\Theta}_{\varepsilon,i}^{k+1} = \frac{n_i}{6\rho}\mathbf{Q}_i\Big(\boldsymbol{\Lambda}_i + \sqrt{\boldsymbol{\Lambda}_i^2 + \frac{12\rho}{n_i}\mathbf{I}}\Big)\mathbf{Q}_i'. \tag{B.18}$$

Step (B.18) is the most computationally intensive task in the algorithm since the runtime of decomposing a $p \times p$ matrix is $\mathcal{O}(p^3)$. Also, note that compared to standard ADMM without smoothing penalty β, (B.18) enforces stronger shrinkage. This is consistent with our motivation for the additional constraint - to smooth the estimator of precision matrix.