

# Graphical lasso for extremes

Phyllis Wan\*

Chen Zhou†

## Abstract

In this paper we estimate the sparse dependence structure in the tail region of a multivariate random vector, potentially of high dimension. The tail dependence is modeled via a graphical model for extremes embedded in the Hüsler-Reiss distribution (Engelke and Hitz, 2020). We propose the *extreme graphical lasso* procedure to estimate the sparsity in the tail dependence, similar to the Gaussian graphical lasso method in high dimensional statistics. We prove its consistency in identifying the graph structure and estimating model parameters. The efficiency and accuracy of the proposed method are illustrated in simulated and real examples.

*Keywords and phrases:* graphical lasso; graphical models; multivariate extreme value statistics; high dimensional statistics; Hüsler-Reiss distribution;

*AMS 2010 Classification:* 62G32; 62H12; 62F12.

## 1 Introduction

Consider a random Gaussian vector with mean zero and covariance matrix  $\Sigma$ . In a Gaussian graphical model, the precision matrix  $\Theta := \Sigma^{-1}$  encodes the conditional dependence structure among the variables – variables  $i$  and  $j$  are conditionally independent given the rest of the variables if and only if  $\Theta_{ij} = 0$  (Lauritzen, 1996).

Given an estimate of the covariance  $\hat{\Sigma}$ , the *graphical lasso* method estimates a sparse  $\Theta$  using  $L_1$ -regularization by

$$\arg \min_{\Theta} \left\{ -\log |\Theta| + \text{tr} \left( \hat{\Sigma} \Theta \right) + \gamma_n \sum_{i \neq j} |\Theta_{ij}| \right\};$$

see, e.g. Yuan and Lin (2007), Banerjee et al. (2008) and Friedman et al. (2008). The advantage of the graphical lasso method is two folds. First, it reveals the conditional dependence among the underlying random variables by producing a sparse estimate of  $\Theta$ . Second, it provides a reliable estimation of  $\Theta$  and  $\Sigma$  in the high dimensional case where classical covariance estimation fails. The theoretical properties of the graphical lasso procedure were investigated in Rothman et al. (2008) and Ravikumar et al. (2011).

---

\*Erasmus University Rotterdam; Econometric Institute, Burg. Oudlaan 50, 3062 PA Rotterdam, the Netherlands; email: wan@ese.eur.nl

†Erasmus University Rotterdam; Econometric Institute, Burg. Oudlaan 50, 3062 PA Rotterdam, the Netherlands; email: zhou@ese.eur.nl

In this paper, we aim to estimate the sparse dependence structure in the *tail region* among high dimensional random variables. With the characterization of tail dependence, one can further conduct statistical risk assessment of extreme (co-)occurrences, such as systemic banking failures (e.g. Zhou (2010)) or compound environmental events (e.g. Coles and Tawn (1991)). Our approach is built on the framework of Engelke and Hitz (2020), which introduces graphical models for extremes by defining the conditional dependence in the tail distribution.

A parametric distribution family that can accommodate sparse graphical models for extremes is the Hüsler-Reiss (HR) distribution (Hüsler and Reiss, 1989). The class of HR distributions describes the non-trivial limiting tail distributions of Gaussian triangular arrays. Similar to Gaussian distribution, its parametrized by bilateral relations. More specifically, a  $d$ -dimensional HR graphical model can be parametrized by a precision matrix  $\Theta \in \mathbb{R}^{d \times d}$ , such that the variables  $i$  and  $j$  are conditionally independent in the extremes given the rest of the variables if and only if  $\Theta_{ij} = 0$  (Engelke and Hitz, 2020, Hentschel et al., 2022).

Unlike the Gaussian case, the precision matrix  $\Theta$  in the HR model is not of full rank. As a consequence, existing statistical inference procedures for estimating  $\Theta$  in a HR model require conditioning on a chosen dimension being above a high threshold. In turn, one can only estimate  $\Theta^{(k)} \in \mathbb{R}^{(d-1) \times (d-1)}$ , the submatrix of  $\Theta$  where the  $k$ -th row and  $k$ -th column are removed (Engelke and Hitz, 2020). Estimating a HR graphical model is therefore challenging when a sparse  $\Theta$  is desired: a sparse estimate of  $\Theta^{(k)}$  does not guarantee sparsity on the omitted  $k$ -th row and column. Hentschel et al. (2022) proposed an estimation procedure for  $\Theta$  using matrix completion when the sparsity structure of  $\Theta$  was known. To date, the only sparse estimation for  $\Theta$  without knowing the sparsity structure ex-ante was proposed by Engelke et al. (2022). They achieved this goal by aggregating sparse estimates of  $\Theta^{(k)}$  for all  $k = 1, \dots, d$  using a majority vote to decide whether or not each entry of  $\Theta$  should be zero. In other words, their estimation procedure requires estimating  $d$  graphical models which can be computationally intensive for large  $d$ .

In this paper, we propose a direct estimate of  $\Theta$  with a built-in option for sparse estimation via  $L_1$ -regularization. We term it the *extreme graphical lasso*. The core idea is as follows. We show that by adding a positive constant  $c$  to each entry of  $\Theta$ , the matrix

$$\Theta^* := \Theta + c\mathbf{1}\mathbf{1}^T$$

is the inverse of a covariance matrix  $\Sigma^*$  which can be estimated consistently from observations. To impose sparsity on the entries of  $\Theta$ , we only need to shrink the off-diagonal entries of  $\Theta^*$  to  $c$ , which can be achieved in the optimization

$$\arg \min_{\Theta^*} \left\{ -\log |\Theta^*| + \text{tr} \left( \hat{\Sigma}^* \Theta^* \right) + \gamma_n \sum_{i \neq j} |\Theta_{ij}^* - c| \right\}.$$

The extreme graphical lasso method requires solving only one optimization problem and therefore is efficient in handling high dimensional situations. In addition, it results in both graph structure identification and parameters estimation simultaneously. The efficiency and accuracy are the main advantages of this novel method.

We provide the finite sample theory and asymptotic theory for the extreme graphical lasso method. In particular, we show a consistent identification of the graph and accurate estimation of the non-sparse parameters in  $\Theta$ . Empirically, we argue that in the high dimensional case, the extreme graphical lasso method can be further simplified by dropping the constant  $c$ , coinciding with the classical graphical lasso algorithm. We apply the extreme graphical lasso to a real data example to illustrate its usefulness in uncovering the underlying dependence structure of extreme events.

The remainder of the paper is structured as follows. The background for HR graphical models is introduced in Section 2. We present our extreme graphical lasso method in Section 3. The non-asymptotic and asymptotic theories are shown in Section 4. Finally, the performance of the method is illustrated in Section 5.

## 1.1 Notation

We will use the following notations. Let  $\mathbf{0}$  and  $\mathbf{1}$  denote vectors whose elements are all 0's and all 1's respectively. For simplicity, with a slight abuse of notation, we may let them denote vectors of different length in different contexts. For the norms for matrices:  $\|\cdot\|_\infty$  is the element-wise  $L_\infty$ -norm, both for vectors and matrices;  $|||\cdot|||_\infty$  is the  $l_\infty$ -operator norm for matrices, i.e. the row-maxima of  $L_1$ -norms applied to each row. We note the following properties of these norms:

- Both  $\|\cdot\|_\infty$  and  $|||\cdot|||_\infty$  are norms.
- For matrix  $A$  and vector  $v$ ,  $\|Av\|_\infty \leq |||A|||_\infty \|v\|_\infty$ .
- For matrices  $A$  and  $B$  with compatible dimensions,  $|||AB|||_\infty \leq |||A|||_\infty |||B|||_\infty$ .

## 2 Hüsler-Reiss graphical models

In this section, we describe the class of HR graphical models and the corresponding statistical inference procedure in existing literature.

### 2.1 Graphical models for extremes

Consider a random vector  $\mathbf{X} = (X_1, \dots, X_d)$ . Denote  $\tilde{X}_k = \frac{1}{1-F_k(X_k)}$ , where  $F_k$  is the marginal distribution function of  $X_k$ . Then  $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_d)$  is a random vector with standard Pareto marginals and summarizes the dependence structure of  $\mathbf{X}$ . Following multivariate extreme value theory, we assume that  $\tilde{\mathbf{X}}$  belongs to the domain of attraction of a multivariate extreme value distribution, i.e. the limit of its component-wise maxima converges to a non-degenerate distribution. Specifically, given i.i.d. copies of  $\tilde{\mathbf{X}}$ ,  $\tilde{\mathbf{X}}^i = (\tilde{X}_1^i, \dots, \tilde{X}_d^i), i \in \mathbb{N}$ , there exists a random vector  $\mathbf{Z} = (Z_1, \dots, Z_d)$  such that

$$P(\mathbf{Z} \leq \mathbf{z}) := \lim_{n \rightarrow \infty} P\left(\max_{i=1, \dots, n} \tilde{X}_1^i \leq nz_1, \dots, \max_{i=1, \dots, n} \tilde{X}_d^i \leq nz_d\right) = G(\mathbf{z}), \quad (2.1)$$

where each marginal distribution of  $G$ ,  $G_k$ , is Fréchet distributed  $G_k(z_k) = \exp(-1/z_k)$ . By writing

$$G(\mathbf{z}) = \exp(-\Lambda(\mathbf{z})),$$

where  $\Lambda(\mathbf{z})$  is shorthand for  $\Lambda([0, \infty)^d \setminus [\mathbf{0}, \mathbf{z}])$ ,  $\Lambda$  is a Radon measure on the cone  $\mathcal{E} = [0, \infty)^d \setminus \{\mathbf{0}\}$ . The measure  $\Lambda$  is known as the exponent measure and characterizes the dependence structure of  $\mathbf{X}$  in the tail region.

The domain of attraction condition (2.1) can be equivalently expressed in terms of threshold exceeding. Consider the exceedances of  $\tilde{\mathbf{X}}$  where its  $L_\infty$ -norm  $\|\tilde{\mathbf{X}}\|_\infty$  is higher than a certain threshold. Then there exists a random vector  $\mathbf{Y}$  such that

$$P(\mathbf{Y} \leq \mathbf{z}) := \lim_{u \rightarrow \infty} P\left(\frac{\tilde{\mathbf{X}}}{u} \leq \mathbf{z} \mid \|\tilde{\mathbf{X}}\|_\infty > u\right) = \frac{\Lambda(\mathbf{z} \wedge \mathbf{1}) - \Lambda(\mathbf{z})}{\Lambda(\mathbf{1})}. \quad (2.2)$$

Here the random vector  $\mathbf{Y}$  is defined with support on the  $L$ -shaped set  $\mathcal{L} = \{\mathbf{x} \in \mathcal{E} : \|\mathbf{x}\|_\infty > 1\}$ . Its distribution is known as a multivariate Pareto distribution.

Engelke and Hitz (2020) proposed the framework of graphical models for extremes, by considering the conditional independence of the threshold exceedance limit  $\mathbf{Y}$  in (2.2). Since  $\mathbf{Y}$  is defined on the  $L$ -shaped set  $\mathcal{L} = \{\mathbf{x} \in \mathcal{E} : \|\mathbf{x}\|_\infty > 1\}$  which is not a product space, the notion of conditional independence is instead defined on the subspace  $\mathcal{L}^k = \{\mathbf{x} \in \mathcal{L} : x_k > 1\}$  for each  $k$ .

Let  $\mathcal{G} = (V, E)$  be a graph defined by a set of nodes  $V = \{1, \dots, d\}$  and a set of undirected edges between pairs of distinct nodes  $E \subset V \times V$ . Define the random vector  $\mathbf{Y}^k \stackrel{d}{=} \mathbf{Y} | Y_k > 1$ . A graphical model for extremes based on graph  $\mathcal{G}$  has a multivariate Pareto distribution  $\mathbf{Y}$  that satisfies

$$\forall k \in \{1, \dots, d\} : Y_i^k \perp\!\!\!\perp Y_j^k | \mathbf{Y}_{\setminus\{i,j\}}^k \Leftrightarrow \{i, j\} \notin E, \quad (2.3)$$

where  $\mathbf{Y}_{\setminus\{i,j\}}^k$  indicates all other dimensions in  $\mathbf{Y}^k$  excluding  $\{i, j\}$ . In short, we denote the conditional independence in extremes as

$$Y_i \perp\!\!\!\perp_e Y_j | \mathbf{Y}_{\setminus\{i,j\}} \Leftrightarrow \{i, j\} \notin E.$$

## 2.2 Hüsler-Reiss graphical models

A  $d$ -dimensional HR model is parametrized by a variogram matrix  $\Gamma \in \mathbb{R}^{d \times d}$ , such that  $\Gamma_{ij} = E(W_i - W_j)^2$  for some centered multivariate Gaussian random vector  $\mathbf{W} = (W_1, \dots, W_d)$ . It is the class of distributions describing the non-trivial tail limiting distribution of Gaussian triangular arrays (Hüsler and Reiss, 1989). Specifically, for any  $k = 1, \dots, d$ , the exponent measure  $\Lambda(\cdot)$  of the HR model admits the density

$$\lambda(\mathbf{y}) = y_k^{-2} \prod_{i \neq k} y_i^{-1} \phi_d(\tilde{\mathbf{y}}_k; \tilde{\Sigma}^{(k)}),$$

where  $\phi_d(\cdot; \tilde{\Sigma}^{(k)})$  is the density of a centered  $d$ -dimensional Gaussian distribution with covariance matrix  $\tilde{\Sigma}^{(k)}$ ,  $\tilde{\mathbf{y}}_k = \{\log(y_i/y_k) + \Gamma_{ik}/2\}_{i=1, \dots, d}$ , and

$$\tilde{\Sigma}^{(k)} = \frac{1}{2} \{\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}\}_{i,j} \in \mathbb{R}^{d \times d}. \quad (2.4)$$

Note that  $\tilde{\Sigma}^{(k)}$  is degenerate on the  $k$ -th row and the  $k$ -th column.

Let  $\Sigma^{(k)}$  be the  $(d-1) \times (d-1)$  matrix constructed by removing the  $k$ -th row and the  $k$ -th column from  $\tilde{\Sigma}^{(k)}$ . For convenience we index the rows and columns of  $\Sigma^{(k)}$  using the original row and column numbers from  $\tilde{\Sigma}^{(k)}$ . Then the following relation holds (Engelke and Hitz, 2020):

$$Y_i \perp\!\!\!\perp_e Y_j | \mathbf{Y}_{\setminus\{i,j\}} \Leftrightarrow \left(\Sigma^{(k)}\right)_{ij}^{-1} = 0, \quad \forall k \neq i, j.$$

Consequently, a zero element in  $\Theta^{(k)} := (\Sigma^{(k)})^{-1}$  corresponds to a non-edge in  $\mathcal{G}$ . In other words,  $\Theta^{(k)}$  serves as the precision matrix for the graph  $\mathcal{G}$  excluding the node  $k$ .

To summarize the graph structure for all dimensions, there exists a precision matrix  $\Theta \in \mathbb{R}^{d \times d}$  such that removing the  $k$ -th column and the  $k$ -th row from  $\Theta$  results in  $\Theta^{(k)}$  (Hentschel et al., 2022). Based on the precision matrix  $\Theta$ , we have that

$$Y_i \perp\!\!\!\perp_e Y_j | \mathbf{Y}_{\setminus\{i,j\}} \Leftrightarrow \Theta_{ij}^{(k)} = 0, i, j \neq k \Leftrightarrow \Theta_{ij} = 0.$$

We can also reconstruct  $\Theta$  given a single  $\Theta^{(k)}$  via

$$\begin{aligned} \Theta_{ij} &= \Theta_{ij}^{(k)}, & \text{if } i, j \neq k; \\ \Theta_{ik} &= -\sum_{l \neq k} \Theta_{il}^{(k)}, & \text{if } i \neq k \text{ and } j = k; \\ \Theta_{kk} &= \sum_{m, l \neq k} \Theta_{ml}^{(k)}, & \text{if } i = k \text{ and } j = k. \end{aligned}$$

Note that  $\Theta \mathbf{1} = \mathbf{0}$ , which implies that  $\Theta$  is not of full rank.

### 2.3 Statistical inference for the HR model

The standard statistical inference for the HR model relies on the following result. Let  $\mathbf{Y}$  be the multivariate Pareto distribution from a HR model with variogram  $\Gamma$ . Engelke et al. (2015) showed that

$$(\log \mathbf{Y}_{-k} - \log(Y_k) \cdot \mathbf{1}) |_{Y_k > 1} \sim N(\Gamma_{\cdot, k}/2, \Sigma^{(k)}). \quad (2.5)$$

Given i.i.d. observations  $\mathbf{X}^i = (X_1^i, \dots, X_d^i)$ ,  $1 \leq i \leq n$  drawn from  $\mathbf{X}$ , an empirical counterpart of  $(\log \mathbf{Y}_{-k} - \log(Y_k) \cdot \mathbf{1}) |_{Y_k > 1}$  can be constructed as follows. Define the transformed observations

$$\hat{X}_k^i = \frac{1}{1 - \hat{F}_k(X_k^i)},$$

where  $\hat{F}_k(x) = \frac{1}{n+1} \sum_{i=1}^n \mathbb{I}\{X_k^i \leq x\}$  is the empirical distribution function based on  $X_k^i$ 's and  $\mathbb{I}$  is the indicator function. Then  $\hat{\mathbf{X}}^i = (\hat{X}_1^i, \dots, \hat{X}_d^i)$  resembles a sample of  $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_d)$  with  $\tilde{X}_k = \frac{1}{1 - F_k(X_k)}$ , albeit not i.i.d.

Consider an intermediate sequence  $k_n$  such that  $k_n \rightarrow \infty$  and  $k_n/n \rightarrow 0$  as  $n \rightarrow \infty$ . Then as  $n \rightarrow \infty$ ,  $\|\hat{\mathbf{X}}^i\|_\infty > \frac{n}{k_n}$  mimicks the condition  $\|\tilde{\mathbf{X}}\|_\infty > u$  with  $u \rightarrow \infty$ . Therefore,

$$\frac{k_n}{n} \hat{\mathbf{X}}^i \left| \|\hat{\mathbf{X}}^i\|_\infty > \frac{n}{k_n} \right.$$

approximately follows the same distribution as  $\mathbf{Y}$ .

Let  $I = \{i : \|\hat{\mathbf{X}}^i\|_\infty > \frac{n}{k_n}\} = \{i_1, \dots, i_m\}$ , where  $m = |I|$ , indicating the index set corresponding to  $\|\hat{\mathbf{X}}^i\|_\infty > \frac{n}{k_n}$ . Denote  $\hat{\mathbf{Y}}_j = \frac{k_n}{n} \hat{\mathbf{X}}^{i_j}$  for all  $j = 1, \dots, m$ . Let

$$\hat{\mathbf{W}}_i^{(k)} = \log \hat{\mathbf{Y}}_{i,-k} - \log(\hat{Y}_{ik}) \cdot \mathbf{1},$$

and  $I_k$  be the index set that  $I_k = \{i : \hat{Y}_{ik} > 1\}$  (note that  $|I_k| = k_n$ ) for each dimension  $k = 1, \dots, d$ . Then  $\Sigma^{(k)}$  can be estimated by

$$\hat{\Sigma}^{(k)} := \frac{1}{k_n} \sum_{i \in I_k} \left( \hat{\mathbf{W}}_i^{(k)} - \frac{1}{k_n} \sum_{i \in I_k} \hat{\mathbf{W}}_i^{(k)} \right) \left( \hat{\mathbf{W}}_i^{(k)} - \frac{1}{k_n} \sum_{i \in I_k} \hat{\mathbf{W}}_i^{(k)} \right)^T, \quad (2.6)$$

which is the sample covariance matrix using  $\hat{\mathbf{W}}_i^{(k)}$  conditional on  $\hat{Y}_{ik} > 1$ .

Theoretically an estimate of  $\Theta$  can be constructed via  $\hat{\Theta}^{(k)} = \left( \hat{\Sigma}^{(k)} \right)^{-1}$ . To achieve sparsity in  $\Theta^{(k)}$ , any sparse inverse covariance matrix estimation technique can be applied here. However, reconstruction of  $\Theta$  from a sparse  $\hat{\Theta}^{(k)}$  does not guarantee sparsity on the omitted  $k$ -th row and column. Engelke et al. (2022) proposed to estimate a sparse  $\hat{\Theta}^{(k)}$  for each  $k$  and then to use a majority vote to decide whether or not each entry of  $\Theta$  should be zero. This approach is shown to be effective in recovering the sparse structure of  $\Theta$ , when the number of dimension is at low or moderate level. For high dimensional case, tuning  $d$  graphical lasso models can be cumbersome.

In the following, we propose a one-step estimation of  $\Theta$ . The advantage is two folds. First, our computation requirement is significantly lower, especially in the case where  $d$  is large. Second, we simultaneously estimate the graph structure and the non-zero elements in  $\Theta$ . Theoretically, we provide concentration bounds for the estimate  $\hat{\Theta}$ , which would otherwise be difficult to recover through the approach of majority vote.

### 3 The extreme graphical lasso

#### 3.1 One-step estimation of $\Theta$

Recall that for a HR distribution parametrized by a variogram  $\Gamma$ , there exists a centered Gaussian random vector  $\mathbf{W}$  such that

$$E(W_i - W_j)^2 = \Gamma_{ij}.$$

Here the choice of  $\mathbf{W}$  is not unique. However, by considering  $\mathbf{W}' = \mathbf{W} - \bar{W} \cdot \mathbf{1}$  with  $\bar{W} = \frac{1}{d} \sum_{i=1}^d W_i$ , for any such  $\mathbf{W}$ , then  $\mathbf{W}'$  is a centered Gaussian random vector with unique covariance matrix

$$\Sigma := -\frac{1}{2} \left( I - \frac{\mathbf{1}\mathbf{1}^T}{d} \right) \Gamma \left( I - \frac{\mathbf{1}\mathbf{1}^T}{d} \right). \quad (3.1)$$

Here  $\Sigma$  is not of full rank since  $\Sigma \mathbf{1} = \mathbf{0}$ . Hentschel et al. (2022) showed that  $\Sigma$  and  $\Theta$  satisfy

$$\lim_{M \rightarrow \infty} (\Sigma + M \mathbf{1}\mathbf{1}^T)^{-1} = \Theta.$$

In the following proposition, we generalize this result to any fixed  $M > 0$ . The proof is postponed to Appendix A.

**Proposition 3.1.** For any  $M > 0$ ,

$$(\Sigma + M\mathbf{1}\mathbf{1}^T)^{-1} = \Theta + \frac{1}{d^2M}\mathbf{1}\mathbf{1}^T.$$

As a result of Proposition 3.1, given any consistent estimator of  $\Sigma$  and  $M > 0$ , we can directly retrieve a consistent estimator of  $\Theta$ .

To estimate  $\Sigma$ , we use the following proposition which shows the link between  $\Sigma^{(k)}$  and  $\Sigma$ . This relationship was noted in Hentschel et al. (2022). For the completeness of this paper, we provide a formal proof in Appendix A.

**Proposition 3.2.** Recall the definition of  $\tilde{\Sigma}^{(k)}$  in (2.4) where  $\tilde{\Sigma}^{(k)} \in \mathbb{R}^{d \times d}$  is the augmented matrix of  $\Sigma^{(k)} \in \mathbb{R}^{(d-1) \times (d-1)}$ . Also recall the definition of  $\Sigma$  in (3.1). We have

$$\frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} = \Sigma + M_{\Sigma} \mathbf{1}\mathbf{1}^T,$$

where

$$M_{\Sigma} = \frac{1}{d^3} \sum_{k=1}^d \mathbf{1}^T \tilde{\Sigma}^{(k)} \mathbf{1}.$$

Based on the estimators for  $\Sigma^{(k)}$ 's in (2.6) and Proposition 3.2, we estimate  $\Sigma$  by

$$S := \frac{1}{d} \sum_{k=1}^d \hat{\Sigma}^{(k)} - \left( \frac{1}{d^3} \sum_{k=1}^d \mathbf{1}^T \hat{\Sigma}^{(k)} \mathbf{1} \right) \mathbf{1}\mathbf{1}^T. \quad (3.2)$$

According to Proposition 3.1, for any fixed  $M > 0$ ,  $\Theta$  can be estimated by

$$\hat{\Theta} := (S + M\mathbf{1}\mathbf{1}^T)^{-1} - \frac{1}{d^2M}\mathbf{1}\mathbf{1}^T. \quad (3.3)$$

### 3.2 Interpretation as aggregated MLE

Given any fixed  $M > 0$ , denote

$$\begin{aligned} \Sigma^* &= \Sigma + M\mathbf{1}\mathbf{1}^T, \\ \Theta^* &= \Theta + \frac{1}{d^2M}\mathbf{1}\mathbf{1}^T, \end{aligned}$$

and

$$S^* = S + M\mathbf{1}\mathbf{1}^T.$$

Then the estimator in (3.3) is equivalent to estimating  $\Theta^*$  by  $(S^*)^{-1}$ , which can also be viewed as the optimizer of the following problem:

$$\arg \min_{\Theta^*} \{-\log |\Theta^*| + \text{tr}(S^* \Theta^*)\}. \quad (3.4)$$

We remark that (3.4) is equivalent to an ‘‘aggregated MLE’’ when considering all partial optimizations in estimating  $\Theta^{(k)}$  as follows.

The pseudo-MLE derived from (2.5) ('pseudo' because the mean of the Gaussian distribution is pre-estimated) is

$$\arg \min_{\Theta^{(k)}} \left\{ -\log |\Theta^{(k)}| + \text{tr} \left( \hat{\Sigma}^{(k)} \Theta^{(k)} \right) \right\}.$$

Combining all  $k$ , one can solve for  $\Theta$  that optimizes

$$\arg \min_{\Theta} \sum_{k=1}^d \left\{ -\log |\Theta^{(k)}| + \text{tr} \left( \hat{\Sigma}^{(k)} \Theta^{(k)} \right) \right\}. \quad (3.5)$$

The following proposition shows that the solution to this aggregated MLE is exactly the same as the estimator  $\hat{\Theta}$  defined in (3.3).

**Proposition 3.3.** *For any fixed  $M > 0$ ,*

$$\frac{1}{d} \sum_{k=1}^d \left\{ -\log |\Theta^{(k)}| + \text{tr} \left( \hat{\Sigma}^{(k)} \Theta^{(k)} \right) \right\} = -\log |\Theta^*| + \text{tr} (S^* \Theta^*) + \log(M) - \frac{\text{tr}(S^*)}{d^2 M}.$$

*Consequently, the optimization problems (3.4) and (3.5) result in the same solution for  $\Theta$ .*

### 3.3 Sparse estimation of $\Theta$

We aim to estimate a sparse  $\Theta$  where some off-diagonal elements are zero. Note that each zero entry of  $\Theta$  corresponds to an entry of  $\Theta^*$  with value  $\frac{1}{d^2 M}$ . We therefore propose the following extreme graphical lasso algorithm by shrinking off-diagonal entries of the matrix  $\Theta^*$  towards  $\frac{1}{d^2 M}$ :

$$\hat{\Theta}^* := \arg \min_{\Theta^*} \left\{ -\log |\Theta^*| + \text{tr} (S^* \Theta^*) + \gamma_n \sum_{i \neq j} \left| \Theta_{ij}^* - \frac{1}{d^2 M} \right| \right\}, \quad (3.6)$$

where  $\gamma_n$  is a suitably chosen penalty parameter. The sparse estimator for  $\Theta$  is then:

$$\hat{\Theta}_{lasso} := \hat{\Theta}^* - \frac{1}{d^2 M} \mathbf{1}\mathbf{1}^T. \quad (3.7)$$

We term this estimation procedure as the *extreme graphical lasso* method.

## 4 Theoretical results

In this section, we establish the finite sample and asymptotic theories for the extreme graphical lasso method in (3.6). The goal is to learn the graphical structure for extremes and estimate the non-zero parameters in  $\Theta$  simultaneously. We start with theoretical results for the estimation of  $\Sigma$  by  $S$ , related to existing theory on the estimation of  $\Gamma$  in Engelke et al. (2022).

### 4.1 Conditions for estimating $\Sigma$

Recall  $S$  as an estimator for  $\Sigma$  in (3.2). We first present the assumptions needed for the finite sample theory of  $S$ . The assumptions are in line with those needed in Theorem 3 in Engelke et al. (2022).

The following condition is needed regarding the tail behavior of  $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_d)$ .

**Condition 4.1** (Assumption 3 in Engelke et al. (2022)). *Assume that all marginal distributions of the original random vector  $\mathbf{X}$ ,  $F_1, \dots, F_d$  are continuous. In addition, there exists  $\xi' > 0$  and  $K' < \infty$  independent of  $d$  such that for all  $J \subset \{1, \dots, d\}$  with  $|J| \in \{2, 3\}$  and  $q \in (0, 1]$ ,*

$$\sup_{x \in [0, 1/q]^2 \times [0, 1]} \left| \frac{1}{q} \mathbb{P} \left( \bigcap_{i \in J} \left\{ \tilde{X}_i > \frac{1}{qx_i} \right\} \right) - \frac{\mathbb{P} \left\{ \bigcap_{i \in J} \tilde{Y}_i > \frac{1}{x_i} \right\}}{\mathbb{P}(Y_1 > 1)} \right| \leq K' q^{\xi'}.$$

Condition 4.1 is a standard second order condition quantifying the speed of convergence of the tail distribution of  $\tilde{\mathbf{X}}$  towards the limiting distribution  $\mathbf{Y}$  on bounded sets. It has been imposed in other asymptotic theories in multivariate extreme value statistics, see e.g. Einmahl et al. (2012) and Engelke and Volgushev (2022).

Next, we assume that the variogram in the HR distribution  $\Gamma$  has bounded entries.

**Condition 4.2** (Bounded entries). *Assume that the variogram  $\Gamma$  satisfies that  $0 < \underline{\lambda} < \inf_{i \neq j} \sqrt{\Gamma_{ij}} \leq \sup_{i \neq j} \sqrt{\Gamma_{ij}} < \bar{\lambda}$ , with  $\underline{\lambda}$  and  $\bar{\lambda}$  independent of  $d$ .*

Condition 4.2 implies the boundedness in the density of the exponent measure, see Assumption 2 in Engelke et al. (2022): this condition is required for establishing concentration bounds for estimators of  $\Gamma$ . In addition, this condition implies that for any pair  $(i, j)$  with  $i \neq j$ ,  $X_i$  and  $X_j$  are asymptotically dependent.

Then we have the following proposition. Its proof is postponed to Appendix B.

**Proposition 4.1.** *Assume that Conditions 4.1 and 4.2 hold. Then for any  $\xi < \xi'$ , there exists positive constants  $C_1, C_2$  and  $C_3$ , depending on  $\xi, \xi', \underline{\lambda}$  and  $\bar{\lambda}$ , independent of  $d$ , such that for any  $\varepsilon \geq \varepsilon_n := C_2 d^3 \exp\{-\frac{C_3 k_n}{(\log n)^8}\}$ ,*

$$\mathbb{P}(\|S - \Sigma\|_\infty > \delta_n) \leq \varepsilon, \tag{4.1}$$

where

$$\delta_n := C_1 \left\{ \left( \frac{k_n}{n} \right)^\xi \left( \log \left( \frac{k_n}{n} \right) \right)^2 + \frac{1 + \sqrt{\frac{1}{C_3} \log(C_2 d^3 / \varepsilon)}}{\sqrt{k_n}} \right\}$$

and  $\|S - \Sigma\|_\infty$  refers to the element-wise maximum error in the estimation.

In addition, assuming that  $(\log n)^4 \sqrt{\frac{\log d}{k_n}} \rightarrow 0$  as  $n \rightarrow \infty$ , we have, as  $n \rightarrow \infty$ ,

$$\|S - \Sigma\|_\infty = O_P \left( \left( \frac{k_n}{n} \right)^\xi \left( \log \left( \frac{k_n}{n} \right) \right)^2 + \sqrt{\frac{\log d}{k_n}} \right).$$

We remark that this theorem does not require a fixed  $d$  and we allow for  $d = d_n \rightarrow \infty$  as  $n \rightarrow \infty$  in the second half of the Proposition. Nevertheless, the condition  $(\log n)^4 \sqrt{\frac{\log d}{k_n}} \rightarrow 0$  as  $n \rightarrow \infty$  provides an upper bound for the diverging speed of  $d_n$  towards infinity. It depends not only on  $n$  but also on the intermediate sequence  $k_n$ .

## 4.2 Conditions for graph identification

Recall that  $(V, E)$  denotes the set of nodes and edges in the graph. Denote  $D = \max_{1 \leq i \leq d} \sum_{j=1}^d \mathbf{1}_{(i,j) \in E}$  as the maximum degree of all nodes. If the dimension  $d = d_n \rightarrow \infty$  as the sample size  $n \rightarrow \infty$ , we can potentially have  $D \rightarrow \infty$  and  $|E| \rightarrow \infty$ . Nevertheless, we always have  $D = O(d)$  and  $|E| = O(d^2)$ .

Here we list the conditions required for learning the graph structure. The first condition concerns the structure of the graph reflected in the matrix  $\Sigma$ .

**Condition 4.3** (Mutual incoherence). *Given  $M > 0$ , define  $\Omega = \Sigma^* \otimes \Sigma^*$  where  $\otimes$  is the Kroneker product. We assume that there exists  $0 < \alpha < 1$  such that*

$$\|\|\Omega_{E^c E}(\Omega_{EE})^{-1}\|\|_\infty < 1 - \alpha,$$

where  $\Omega_{EE} \in \mathbb{R}^{|E| \times |E|}$  is the submatrix  $(\Omega_{(i,j),(k,l)})_{(i,j) \in E, (k,l) \in E}$  and  $\Omega_{E^c E}$  is defined similarly.

Note that the *mutual incoherence* condition (sometimes referred to as the irrerepresentability condition) is comparable with Assumption 1 in Ravikumar et al. (2011). Such a condition is often needed for theory regarding lasso-type penalization algorithms. Given a graph structure and a matrix  $\Sigma$ , the validity of our mutual incoherence condition depends on the choice of  $M$ . We illustrate this in Section 5.1 with two examples.

The next condition concerns the tuning parameter  $\gamma_n$ . In order to identify the graphical structure precisely, the tuning parameter should be neither too high nor too low. A low  $\gamma_n$  will result in non-edges not being penalized to zero while a high  $\gamma_n$  will penalize true edges to zero. Therefore, we need both an upper and a lower bound for  $\gamma_n$ . The following condition is formulated for a fixed constant  $\epsilon > 0$  to be specified in Condition 4.5.

**Condition 4.4.** *Assume that the tuning parameter  $\gamma_n$  satisfies  $\underline{C}_\gamma(\delta_n) \leq \gamma_n \leq \overline{C}_\gamma$ , where*

$$\overline{C}_\gamma := \frac{(1 - \epsilon)\alpha(1 - \alpha)}{D\|\|\Sigma^*\|\|_\infty\|\|(\Omega_{EE})^{-1}\|\|_\infty [(1 - \epsilon)\alpha + \|\|\Sigma^*\|\|_\infty^2\|\|(\Omega_{EE})^{-1}\|\|_\infty]}, \quad (4.2)$$

$$\underline{C}_\gamma(\delta_n) := \frac{1 - \alpha}{\epsilon\alpha} \cdot \delta_n, \quad (4.3)$$

where  $\delta_n$  quantifies the estimation error for  $S - \Sigma$  as in Proposition 4.1.

The upper bound  $\overline{C}_\gamma$  is a constant related to the parameters of the graphical model only. The lower bound is a linear function of  $\delta_n$ . In the asymptotic setup, it tends to 0 as  $n \rightarrow \infty$ .

The last condition ensures that the above bounds can be achieved, that is  $\underline{C}_\gamma \leq \overline{C}_\gamma$ . Note that this condition will be satisfied when  $n$  is sufficiently large.

**Condition 4.5.** *There exists an  $\epsilon > 0$  such that*

$$\delta_n \leq \frac{(1 - \epsilon)\epsilon\alpha^2}{D\|\|\Sigma^*\|\|_\infty\|\|\Omega_{EE}^{-1}\|\|_\infty [(1 - \epsilon)\alpha + \|\|\Sigma^*\|\|_\infty^2\|\|\Omega_{EE}^{-1}\|\|_\infty]},$$

where  $\delta_n$  quantifies the estimation error in  $S - \Sigma$  as in Proposition 4.1.

### 4.3 Main theorem

We first present the concentration bounds for  $\hat{\Theta}_{lasso}$  for fixed  $n$ . The proof is shown in Appendix C.

**Theorem 4.2.** *Assume that Conditions 4.1–4.3 holds. For some  $\varepsilon \geq \varepsilon_n$  with  $\varepsilon_n$  defined in Proposition 4.1, further assume that Conditions 4.4–4.5 hold with  $\delta_n$  defined in Proposition 4.1. Then on an event  $A_\varepsilon$  with  $\mathbb{P}(A_\varepsilon) > 1 - \varepsilon$ , the extreme graphical lasso algorithm specified in (3.6) and (3.7) has a unique solution  $\hat{\Theta}^*$  and  $\hat{\Theta}_{lasso}$ . In addition, for this solution, denote the estimated edges as  $\hat{E} := \{(i, j) : \hat{\Theta}_{lasso,ij} \neq 0\}$ . We have that on  $A_\varepsilon$ ,*

$$\hat{E} \subset E$$

and

$$\|\hat{\Theta}_{lasso} - \Theta\|_\infty \leq \frac{\|(\Omega_{EE})^{-1}\|_\infty}{1 - \alpha} \cdot \gamma_n =: r_n. \quad (4.4)$$

In particular, if  $\min\{\|\Theta_{ij}\|; (i, j) \in E, i \neq j\} > r_n$ , then we have that  $\hat{E} = E$  on  $A_\varepsilon$ .

Next, we present the asymptotic theory when  $n \rightarrow \infty$ . Notice that by assuming  $k_n/n \rightarrow 0$  and  $(\log n)^4 \sqrt{\frac{\log d}{k_n}} \rightarrow 0$  as  $n \rightarrow \infty$ , we have  $\delta_n \rightarrow 0$ . Then for any  $\epsilon > 0$ , Condition 4.5 is satisfied for sufficiently large  $n$ . To achieve the lowest estimation error, we choose the lowest possible tuning parameter  $\gamma_n = \underline{C}_\gamma(\delta_n)$  as in (4.3). This implies that both  $\gamma_n = O(\delta_n)$  and  $r_n = O(\delta_n)$  as  $n \rightarrow \infty$ . The following asymptotic result follows immediately from Theorem 4.2.

**Theorem 4.3.** *Assume that Conditions 4.1–4.3 holds. Choose the tuning parameter  $\gamma_n = \underline{C}_\gamma(\delta_n)$  as in (4.3). Assume that  $\min\{\|\Theta_{ij}\|; (i, j) \in E, i \neq j\} > r > 0$  with some constant  $r$  independent of  $d$ . Then as  $n \rightarrow \infty$ , with probability tending to 1, the solution for  $\hat{\Theta}^*$  and  $\hat{\Theta}_{lasso}$  is unique. In addition,*

$$\Pr(\hat{E} = E) \rightarrow 1 \quad \text{and} \quad \|\hat{\Theta}^* - \Theta^*\|_\infty = O_P \left\{ \left( \frac{k}{n} \right)^\xi \left( \log \frac{n}{k} \right)^2 + \sqrt{\frac{\log d}{k}} \right\}.$$

## 5 Simulations and a real data example

In this section, we demonstrate the performance of the extreme graphical lasso method in both the low dimensional ( $d = 4$ ) and high dimensional ( $d = 20$  and  $d = 200$ ) cases. In addition, we show a data example of river discharge data, also used in Engelke and Hitz (2020).

### 5.1 Simulations in low dimensional cases ( $d = 4$ )

For  $d = 4$ , we investigate two theoretical examples. In particular, we focus on the Mutual Incoherence condition

$$\|(\Omega_{E^c E}(\Omega_{EE})^{-1})\|_\infty < 1,$$

and demonstrate how the validity of this condition depends on the choice of  $M$ .

Figure 1 shows the graph structures for the two examples, the star graph and the diamond graph. The Mutual Incoherence condition for these two graphs in the classical graphical lasso setting was studied in Ravikumar et al. (2011).

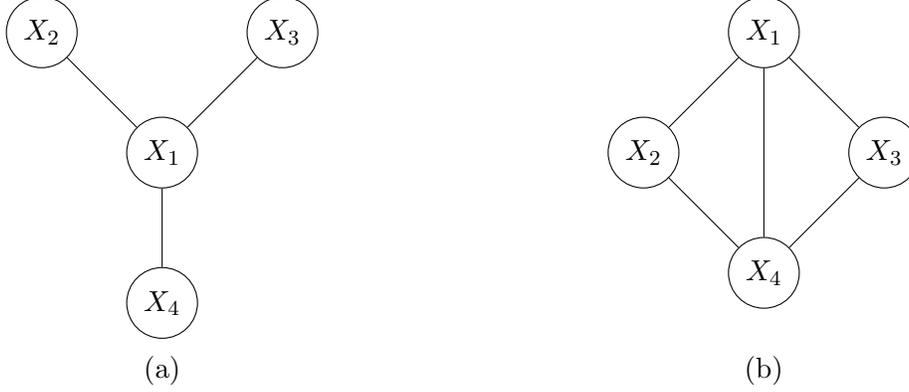


Figure 1: (a) Star graph; (b) Diamond graph.

### 5.1.1 Star graph

Notice that the precision matrix  $\Theta$  in the HR model satisfies the constraint  $\Theta \mathbf{1} = \mathbf{0}$ , which leaves limited options for  $\Theta$  given a certain sparsity structure. We consider the following parameterization

$$\Theta = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

which reflects the Star graph in Figure 1(a).

We plot the values of  $|||\Omega_{E^c E}(\Omega_{EE})^{-1}|||_{\infty}$  against the values of  $M$  on the left panel of Figure 2. The figure shows that the Mutual Incoherence condition is satisfied when  $M \in (0, 0.2768]$ .

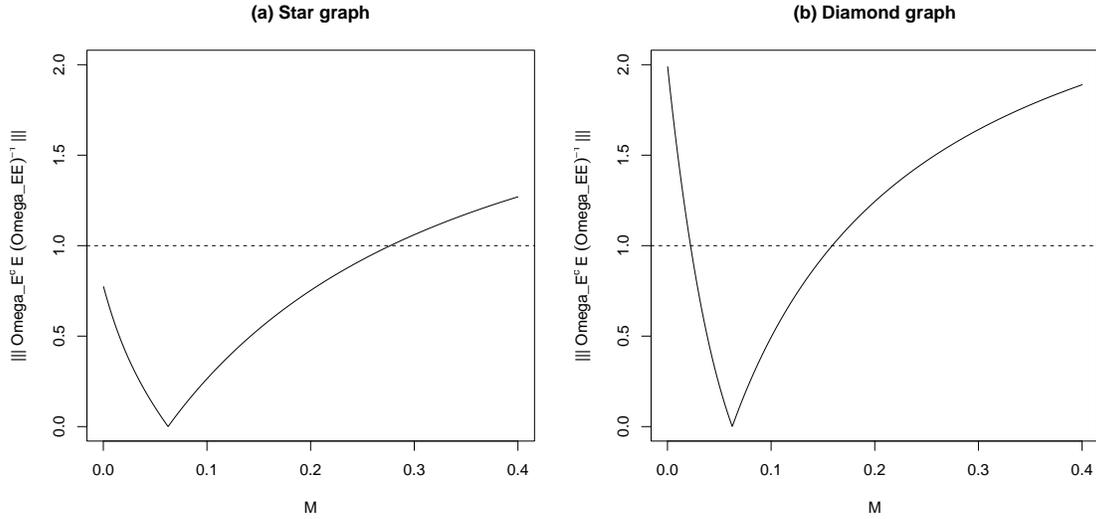


Figure 2: The curves of  $\|\Omega_{E^c E}(\Omega_{EE})^{-1}\|_\infty$  versus  $M$  for the star graph and the diamond graph.

We simulate 100 samples with various sample sizes  $n$ : ranging from  $10^3$  to  $10^6$ . For each sample we simulate data from the following multivariate Pareto distribution

$$\mathbf{X} = Y \exp\{\mathbf{W} - \sigma^2(\mathbf{W})/2\},$$

where  $\mathbf{W}$  follows a zero-mean Gaussian distribution with a covariance matrix  $\Sigma$  calculated based on Proposition 3.1, and  $Y$  follows a standard Pareto distribution and is independent of  $\mathbf{W}$ . Here  $\sigma^2(\mathbf{W})$  is the diagonal vector of  $\Sigma$ . Note that the choice of  $M$  here is irrelevant to the calculation of  $\Sigma$ . The multivariate Pareto distribution used in the simulation is in the domain of attraction of a HR model with precision matrix  $\Theta$ .

We apply the extreme graphical lasso method to estimate the graphical structure of  $\Theta$ . More specifically, for the estimator  $\hat{\Sigma}$ , we use  $k_n = 0.05n$ . For the extreme graphical lasso, we choose  $M = 0.25$  which ensures the Mutual Incoherence condition and a penalty parameter  $\gamma_n = 0.2$ . Here the optimization problem (3.6) is convex and can be solved with a block coordinate descent algorithm similar to Mazumder and Hastie (2012), see Appendix D.

After obtaining  $\hat{\Theta}$  we further consider a thresholding by 0.01: if an estimated off-diagonal element has an absolute value less or equal to  $10^{-2}$ , it will be set to zero. The last step is purely for computational reason. For each simulated sample, we consider it as a “success” if the estimated graph coincides with the true graph. The left panel of Figure 3 shows the “success rates” in 100 simulated samples versus the sample sizes.

We observe that the success rate of the extreme graphical lasso method approaches 100% as  $n$  increases, indicating that the graph can be consistently identified. Note that here  $k_n$  is the effective sample size and corresponds to 5% of  $n$ , e.g.  $k_n = 500$  corresponding to  $n = 10000$ .

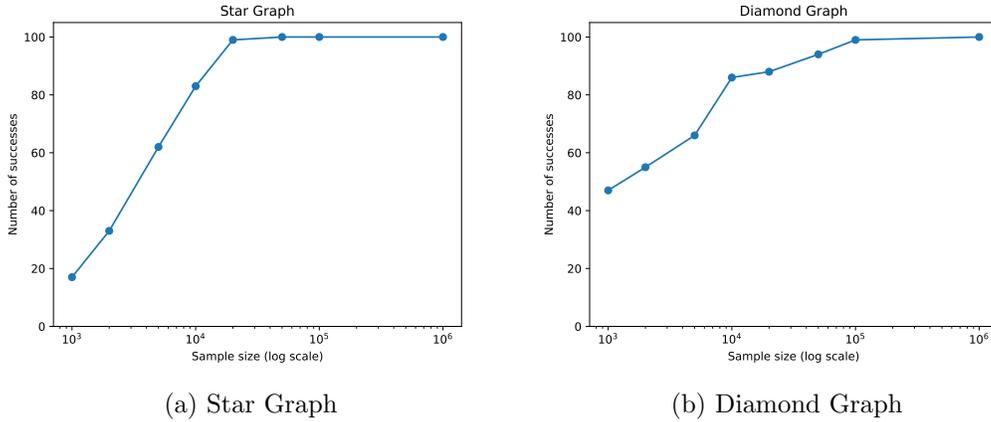


Figure 3: The success rates versus sample sizes based on 100 samples

### 5.1.2 Diamond graph

We now consider the diamond graph in Figure 1(b) corresponding to the following precision matrix

$$\Theta = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}.$$

Again we plot the values of  $\|\Omega_{E^c E}(\Omega_{EE})^{-1}\|_\infty$  against the values of  $M$  on the right panel of Figure 2. The figure shows that the Mutual Incoherence condition is satisfied when  $M \in [0.0224, 0.1588]$ .

We conduct simulations for the diamond graph with the same setup as in Section 5.1.1. In the extreme graphical lasso estimation, we choose  $M = 0.15$  and  $\gamma_n = 0.1$ . The results for the success rate are shown on the right panel of Figure 3. Again the graph structure can be consistently identified for sample sizes starting from  $n = 5000$  corresponding to an effect sample size  $k_n = 250$ . Compared to identifying the Star graph, identifying the Diamond graph is relatively easier.

## 5.2 Simulations in high dimensional cases ( $d = 20$ and $d = 200$ )

We demonstrate the performance of the extreme graphical lasso method in higher dimensional situations.

If the dimension  $d$  is high, the extreme graphical lasso method can be simplified to a standard graphical lasso method as follows. Recall the extreme graphical lasso procedure in (3.6) and (3.7). In the optimization step, we shrink off-diagonal entries of the matrix  $\Theta^*$  towards  $\frac{1}{d^2 M}$ . When  $d$  is large, the term  $\frac{1}{d^2 M}$  is close to zero. Therefore, we can replace the optimization step (3.6) by

$$\hat{\Theta}_{mlasso} := \arg \min_{\Theta^*} \left\{ -\log |\Theta^*| + \text{tr}(S^* \Theta^*) + \gamma_n \sum_{i \neq j} |\Theta_{ij}^*| \right\},$$

potentially with a different penalization parameter  $\gamma_n$ . This practical proposal performs the classical graphical lasso procedure only once. In the estimation, we use the R package *glasso* in Friedman et al. (2008) for the optimization. We term it as the *modified extreme graphical lasso* method.

We first demonstrate that the modified extreme graphical lasso procedure results in similar graphs as the extreme graphical lasso method for  $d = 20$ . For that purpose, we simulate observations following a multivariate Pareto distribution with  $d = 20$ . The multivariate Pareto distribution is in the domain of attraction of a HR model with a specific precision matrix governed by a graph. The true graph is presimulated by a preferential attach model as in Albert and Barabási (2002), see Figure 4(a). Given the graph, the simulation of the observations is achieved by using the R package *graphicalExtremes* in Engelke and Hitz (2020). We simulate 100 samples with sample size 5000.

Across the 100 samples, we apply the extreme graphical lasso method to estimate the graph structure with  $k_n = 250$ ,  $M = 1$  and  $\gamma_n = 2.4$ . In the extreme graphical lasso method, we shrink the off-diagonal elements of  $\hat{\Theta}^*$  towards  $c = 1/(d^2M)$ . For each pair of nodes, we count the number of samples for which an edge is identified. Figure 4(b) shows the aggregation of 100 estimated graphs, where the thickness of each edge reflects the proportion of times an edge is identified.

Next, we apply the modified extreme graphical lasso method based on the same estimated  $\Sigma$ . Here we use  $\gamma_n = 1.2$ . The aggregation of 100 estimated graphs is shown in Figure 4(c).

The two graphs in the panels (b) and (c) of Figure 4 are virtually the same, with the modified procedure identifying slightly more wrong edges. Both graphs are comparable with the true graph in panel (a), indicating the applicability of both procedures.

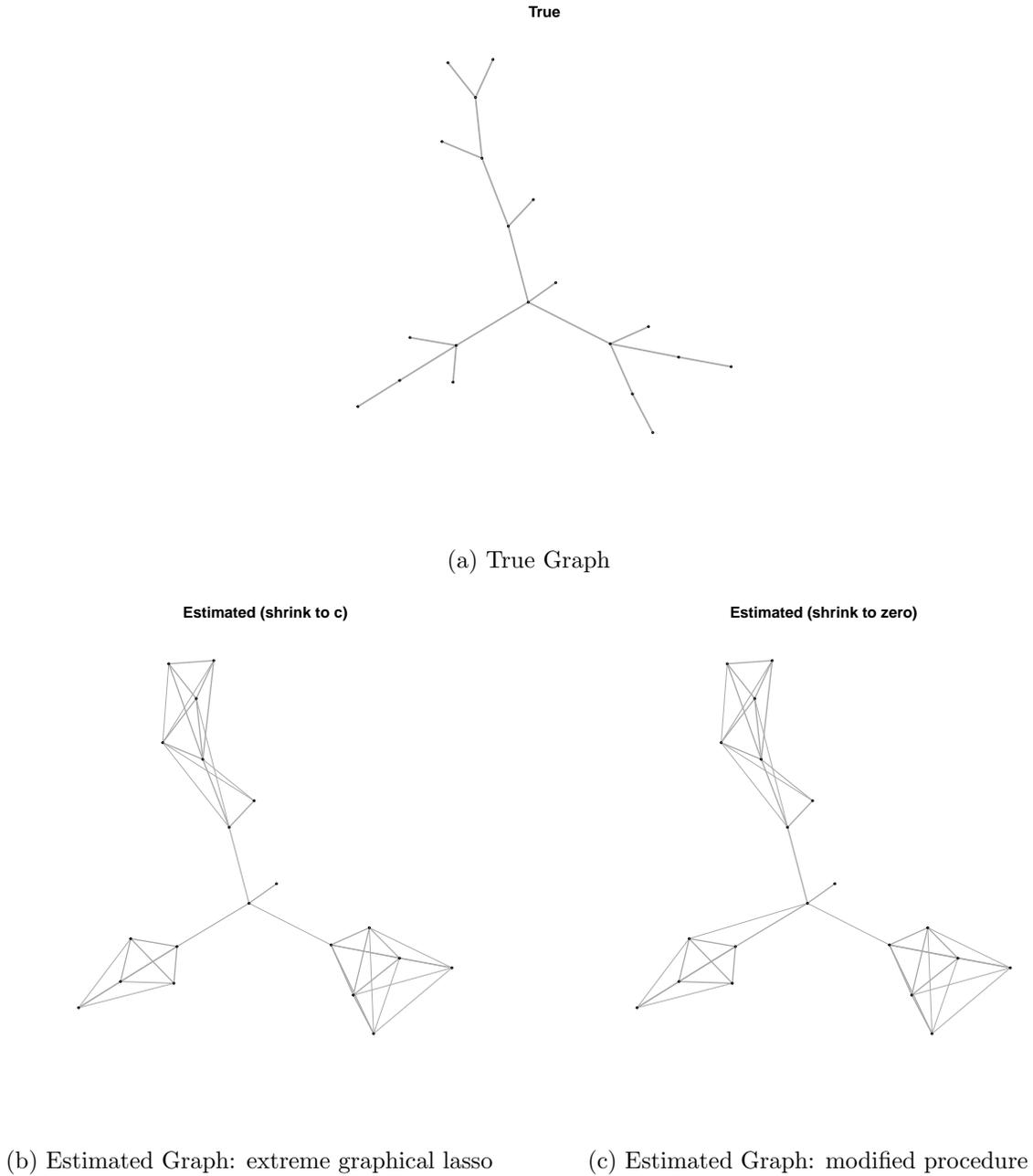


Figure 4: The true and estimated graphs for  $d = 20$  (100 samples)

Now we demonstrate the efficiency of the modified extreme graphical lasso in a very high dimensional case  $d = 200$ . We perform a simulation for  $d = 200$  with  $n = 100000$ . The simulation setup is similar to the  $d = 20$  case, while in the estimation we use the modified extreme graphical lasso procedure with  $k_n = 5000$  and  $\gamma_n = 1.9$ . Here the penalization parameter is chosen at the highest level for which the estimated graph is connected. The true graph and the aggregated graph from 10 simulations are shown in Figure 5.

We particularly focus on the efficiency of the algorithm.<sup>1</sup> The average time for running each of the 10 iteration is 3.23 min. Per iteration, the majority time is devoted to simulating the data (0.378 min) and estimating the  $\Sigma$  matrix (2.848 min), while the time to perform the modified extreme graphical lasso method is only 0.0037 min, about 0.22 second. Considering that this is for  $d = 200$ , the modified extreme graphical lasso method has great potential for handling even higher dimensional cases.

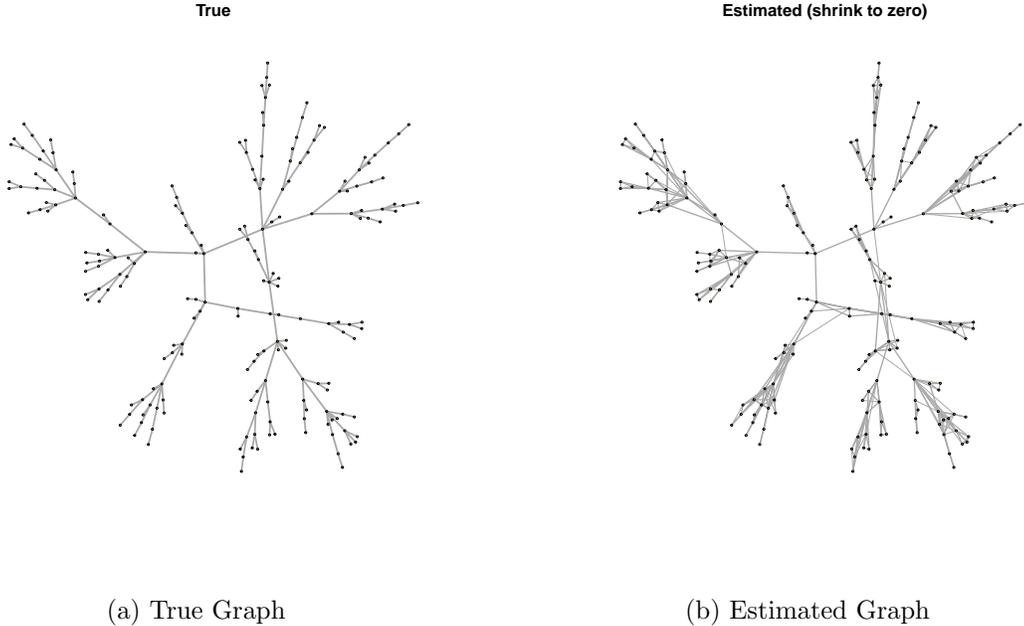


Figure 5: The true and estimated graphs for  $d = 200$  (10 samples)

### 5.3 Real data example: river discharge

We apply the modified extreme graphical lasso method to the river discharge data in the upper Danube basin. This dataset was first analyzed in Asadi et al. (2015) and subsequently studied in Engelke and Hitz (2020). The data contain river discharge at  $d = 31$  stations with a sample size 428 after declustering. The physical locations of the stations and the altitude of the area are shown in Figure 6.<sup>2</sup> We refer interested readers to Asadi et al. (2015) for more information about the dataset and the declustering procedure. We use  $k_n = 64$  in the estimation, and vary the penalization parameter  $\gamma_n$  to obtain different estimated graphs.

In Figure 7, the top left panel shows the physical connection of the stations following the Danube basin. The other three panels in the figure show different estimated graphs using different values of  $\gamma_n$ . With increasing  $\gamma_n$  the estimated graph contains fewer edges and eventually turns to disconnected graphs. Nevertheless, such a graph will be useful for practitioners to understand the

<sup>1</sup>This simulation is run on a Dell XPS 9320 laptop, with 16 cores (i7-1260P) and 32GB memory. Operating system: Ubuntu 22.04.2. R version: 4.3.1.

<sup>2</sup>We are grateful to Sebastian Engelke for providing the figure.

most related extreme river discharge across the stations.

For instance, we observe that the extreme river discharge at Station 1 is more related to the stream from Station 13, instead of the other stream from Station 2. The stream from Station 13 is the downstream of the river Salzach, originally starting in the Alps, while the stream from Station 2 is the Danube river flowing over a plain area. Similarly, the extreme river discharge at Station 2 is more related to the main stream from Station 3 (the Danube river) instead of the branch from Station 14 (the Isar river).

To conclude, by applying the (modified) extreme graphical lasso method to the Danube river discharge data, we obtain insights regarding the interrelations of extreme river discharges at a large number of stations spanning in a large spatial area.

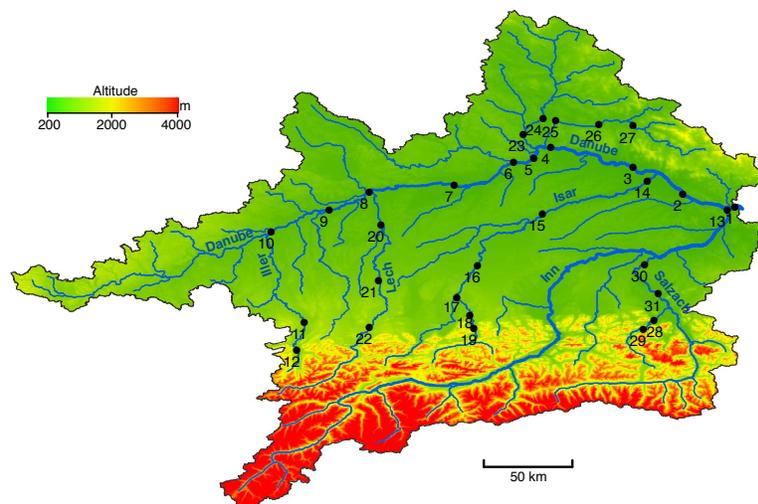


Figure 6: The river map of the upper Danube basin. Courtesy of Asadi et al. (2015).

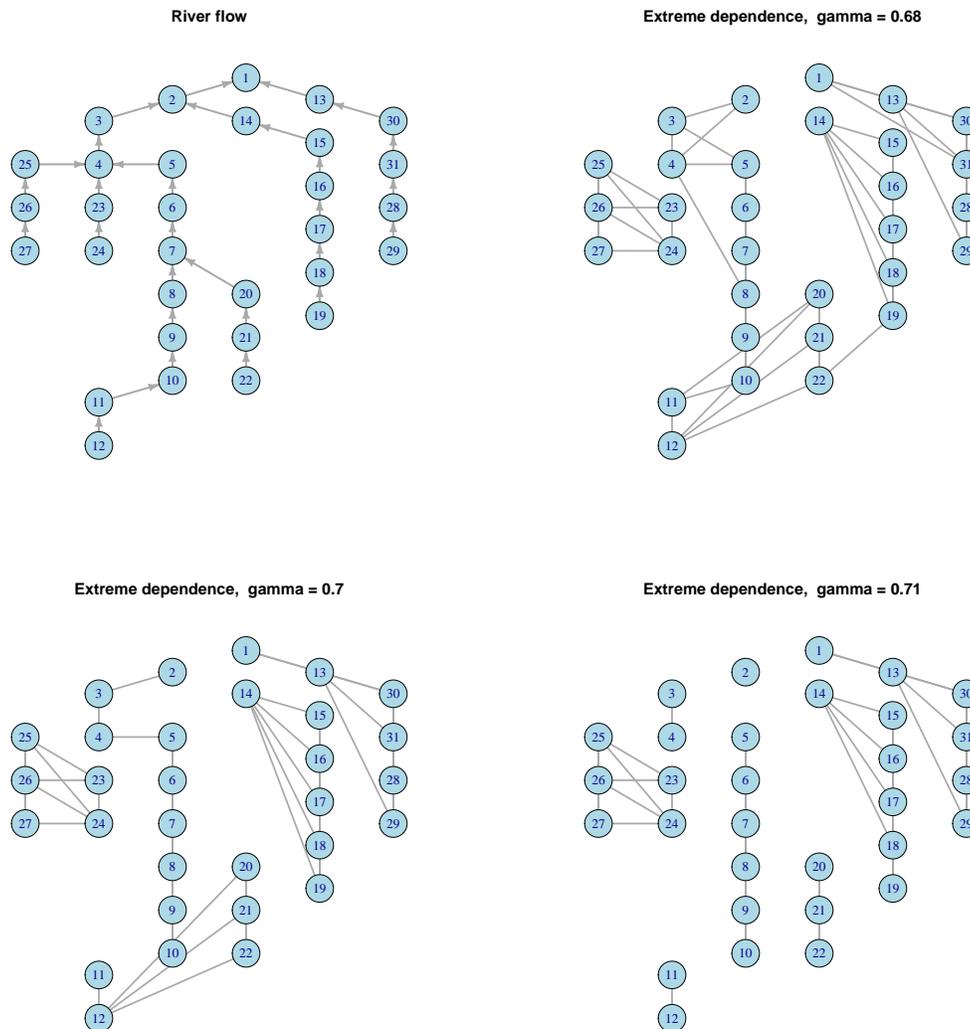


Figure 7: The physical flow connection and estimated graphs for the Danube river discharge data.

## References

- Albert, R. and A.-L. Barabási (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* 74(1), 47.
- Asadi, P., A. C. Davison, and S. Engelke (2015). Extremes on river networks. *Annals of Applied Statistics* 9(4), 2023–2050.
- Banerjee, O., L. El Ghaoui, and A. d’Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research* 9, 485–516.

- Coles, S. G. and J. A. Tawn (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 53, 377–392.
- Einmahl, J., A. Krajina, and J. Segers (2012). An M-estimator for tail dependence in arbitrary dimensions. *Annals of Statistics* 40(3), 1764–1793.
- Engelke, S. and A. Hitz (2020). Graphical models for extremes (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 871–932.
- Engelke, S., M. Lalancette, and S. Volgushev (2022). Learning extremal graphical structures in high dimensions. *arXiv preprint arXiv:2111.00840*.
- Engelke, S., A. Malinowski, Z. Kabluchko, and M. Schlather (2015). Estimation of Hüsler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 77(1), 239–265.
- Engelke, S. and S. Volgushev (2022). Structure learning for extremal tree models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(5), 2055–2087.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441.
- Hentschel, M., S. Engelke, and J. Segers (2022). Statistical Inference for Hüsler–Reiss Graphical Models Through Matrix Completions. *arXiv preprint arXiv:2210.14292*.
- Hüsler, J. and R.-D. Reiss (1989). Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters* 7, 283–286.
- Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.
- Mazumder, R. and T. Hastie (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics* 6, 2125.
- Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* 5, 935–980.
- Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2, 494–515.
- Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* 94(1), 19–35.
- Zhou, C. (2010). Are banks too big to fail? measuring systemic importance of financial institutions. *International Journal of Central Banking* 6(34), 205–250.

## A Proof of propositions in Sections 3

We first prove Proposition 3.2 and then use the result to prove Proposition 3.1.

*Proof of Proposition 3.2.* Note that the  $(i, j)$ th element of  $\frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)}$  is equal to

$$\begin{aligned} \left( \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} \right)_{i,j} &= \frac{1}{2} \frac{1}{d} \sum_{k=1}^d (\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}) \\ &= -\frac{1}{2} \left( \Gamma_{ij} - \frac{1}{d} \sum_{k=1}^d \Gamma_{ik} - \frac{1}{d} \sum_{k=1}^d \Gamma_{jk} \right). \end{aligned}$$

Hence

$$\begin{aligned} \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} &= -\frac{1}{2} \left( \Gamma - \frac{1}{d} \Gamma \mathbf{1} \mathbf{1}^T - \frac{1}{d} \mathbf{1} \mathbf{1}^T \Gamma \right) \\ &= -\frac{1}{2} \left( \Gamma - \frac{1}{d} \Gamma \mathbf{1} \mathbf{1}^T - \frac{1}{d} \mathbf{1} \mathbf{1}^T \Gamma + \frac{1}{d^2} \mathbf{1} \mathbf{1}^T \Gamma \mathbf{1} \mathbf{1}^T \right) + \frac{1}{2d^2} \mathbf{1} \mathbf{1}^T \Gamma \mathbf{1} \mathbf{1}^T \\ &= -\frac{1}{2} \left( I - \frac{\mathbf{1}^T \mathbf{1}}{d} \right) \Gamma \left( I - \frac{\mathbf{1}^T \mathbf{1}}{d} \right) + \left( \frac{\mathbf{1}^T \Gamma \mathbf{1}}{d^2} \right) \cdot \mathbf{1} \mathbf{1}^T \\ &= \Sigma + \left( \frac{\mathbf{1}^T \Gamma \mathbf{1}}{d^2} \right) \cdot \mathbf{1} \mathbf{1}^T. \end{aligned}$$

Summing up the elements of the matrices on both sides, we have

$$\begin{aligned} \mathbf{1}^T \left( \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} \right) \mathbf{1} &= \mathbf{1}^T \Sigma \mathbf{1} + \mathbf{1}^T \left( \left( \frac{\mathbf{1}^T \Gamma \mathbf{1}}{d^2} \right) \cdot \mathbf{1} \mathbf{1}^T \right) \mathbf{1} \\ &= 0 + \mathbf{1}^T \Gamma \mathbf{1}. \end{aligned}$$

Plugging in the value for  $\mathbf{1}^T \Gamma \mathbf{1}$  back into the previous equation, we obtain that

$$\frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} = \Sigma + \frac{1}{d^2} \mathbf{1}^T \left( \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} \right) \mathbf{1} \cdot \mathbf{1} \mathbf{1}^T = \Sigma + \frac{1}{d^3} \left( \sum_{k=1}^d \mathbf{1}^T \tilde{\Sigma}^{(k)} \mathbf{1} \right) \cdot \mathbf{1} \mathbf{1}^T.$$

□

*Proof of Proposition 3.1.* Using the property  $\Sigma \mathbf{1} = \mathbf{0}$  and  $\Theta \mathbf{1} = \mathbf{0}$  and the fact that  $\Sigma$  and  $\Theta$  are both symmetric, we get

$$\begin{aligned} &(\Sigma + M \mathbf{1} \mathbf{1}^T) \cdot \left( \Theta + \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T \right) \\ &= \Sigma \cdot \Theta + M \mathbf{1} \mathbf{1}^T \cdot \Theta + \Sigma \cdot \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T + M \mathbf{1} \mathbf{1}^T \cdot \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T \\ &= \Sigma \cdot \Theta + M \mathbf{1} \cdot (\Theta \mathbf{1})^T + \frac{1}{d^2 M} \cdot \Sigma \mathbf{1} \cdot \mathbf{1}^T + \left( M \cdot \frac{1}{d^2 M} \right) \cdot \mathbf{1} \mathbf{1}^T \cdot \mathbf{1} \mathbf{1}^T \\ &= \Sigma \cdot \Theta + \frac{1}{d} \cdot \mathbf{1} \mathbf{1}^T \end{aligned}$$

From Proposition 3.2, we can write the first term as

$$\begin{aligned}
\Sigma \cdot \Theta &= \left( \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} - M_{\Sigma} \mathbf{1} \mathbf{1}^T \right) \cdot \Theta \\
&= \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} \cdot \Theta - M_{\Sigma} \mathbf{1} \mathbf{1}^T \cdot \Theta \\
&= \frac{1}{d} \sum_{k=1}^d \tilde{\Sigma}^{(k)} \cdot \Theta \\
&= I_{d \times d} - \frac{1}{d} \cdot \mathbf{1} \mathbf{1}^T.
\end{aligned}$$

Therefore

$$(\Sigma + M \mathbf{1} \mathbf{1}^T) \cdot \left( \Theta + \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T \right) = I_{d \times d}.$$

□

In order to prove Proposition 3.3, we make use of the following lemma.

**Lemma A.1.** *For any  $k$ ,*

$$|\Theta^*| = \frac{1}{M} |\Theta^{(k)}|.$$

*Proof of Lemma A.1.* We will first show that for  $k \neq k'$ ,

$$|\Theta^{(k)}| = |\Theta^{(k')}|.$$

Note that

$$\Theta^{(k)} = A_k^T \Theta A_k,$$

where  $A_k \in \mathbb{R}^{d \times (d-1)}$ ,  $A_k[:, k] = \mathbf{0}$  and  $A_k[:, -k] = I_{(d-1) \times (d-1)}$ . In other words,  $A_k$  takes the  $(d-1) \times (d-1)$  identity matrix and insert an extra  $k$ -th row with zero entries. On the other hand, we have

$$\Theta = B_k \Theta^{(k)} B_k^T,$$

where  $B_k \in \mathbb{R}^{d \times (d-1)}$ ,  $B_k[:, k] = -\mathbf{1}$  and  $B_k[:, -k] = I_{(d-1) \times (d-1)}$ . In other words,  $B_k$  takes the  $(d-1) \times (d-1)$  identity matrix and insert an extra  $k$ -th row with  $-1$  entries. Therefore we have

$$\Theta^{(k')} = A_{k'}^T B_k \Theta^{(k)} B_k^T A_{k'}$$

and

$$|\Theta^{(k')}| = |A_{k'}^T B_k| \cdot |\Theta^{(k)}| \cdot |B_k^T A_{k'}|.$$

Now we claim that  $|A_{k'}^T B_k| = 1$  for any  $k, k'$ . We have

$$A_{k'}^T B_k = \sum_{h=1}^d A_{k'}[:, h]^T B_k[h, ]$$

$$\begin{aligned}
&= \sum_{h \neq k'} A_{k'}[h]^T B_k[h,] + A_{k'}[k']^T B_k[k',] \\
&= A_{k'}[-k']^T B_k[-k',] + A_{k'}[k']^T B_k[k',] \\
&= I_{(d-1) \times (d-1)} B_k[-k',] + \mathbf{0}^T B_k[k',] \\
&= B_k[-k',].
\end{aligned}$$

For example, assume that  $k = 1$  and  $k' = d$ , then

$$B_1[-d,] = \begin{pmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

and it is easy to see that  $|B_1[-d,]| = 1$ .

It can be shown by the calculation of determinant that  $|B_k[-k',]| = 1$  for any  $k \neq k'$ . Therefore

$$|A_{k'}^T B_k| = 1, \quad k \neq k',$$

and

$$|\Theta^{(k)}| = |\Theta^{(k')}|, \quad k \neq k'.$$

Now to show that  $|\Theta^*| = \frac{1}{M} |\Theta^{(k)}|$ , it suffices to prove it for one value of  $k$ . We will show it for  $k = d$ .

Note that we have

$$\frac{1}{M} |\Theta^{(d)}| = \left| \begin{pmatrix} \Theta^{(d)} & \mathbf{0} \\ \mathbf{0}^T & \frac{1}{M} \end{pmatrix} \right|.$$

We establish the following transformation

$$\begin{aligned}
&\begin{pmatrix} I & \mathbf{0} \\ -\mathbf{1}^T & 1 \end{pmatrix} \begin{pmatrix} I & \frac{1}{d} \mathbf{1} \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} \Theta^{(d)} & \mathbf{0} \\ \mathbf{0}^T & \frac{1}{M} \end{pmatrix} \begin{pmatrix} I & \mathbf{0} \\ \frac{1}{d} \mathbf{1}^T & 1 \end{pmatrix} \begin{pmatrix} I & -\mathbf{1} \\ \mathbf{0}^T & 1 \end{pmatrix} \\
&= \begin{pmatrix} I & \mathbf{0} \\ -\mathbf{1}^T & 1 \end{pmatrix} \begin{pmatrix} \Theta^{(d)} + \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T & \frac{1}{dM} \mathbf{1} \\ \frac{1}{dM} \mathbf{1}^T & \frac{1}{M} \end{pmatrix} \begin{pmatrix} I & -\mathbf{1} \\ \mathbf{0}^T & 1 \end{pmatrix} \\
&= \begin{pmatrix} \Theta^{(d)} + \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T & -\Theta^{(d)} \mathbf{1} + \frac{1}{d^2 M} \mathbf{1} \\ -\mathbf{1}^T \Theta^{(d)} + \frac{1}{d^2 M} \mathbf{1}^T & \mathbf{1}^T \Theta^{(d)} \mathbf{1} + \frac{1}{d^2 M} \end{pmatrix} \\
&= \Theta + \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T \\
&= \Theta^*.
\end{aligned}$$

Since

$$\left| \begin{pmatrix} I & \mathbf{0} \\ -\mathbf{1}^T & 1 \end{pmatrix} \right| = \left| \begin{pmatrix} I & \mathbf{0} \\ \frac{1}{d} \mathbf{1}^T & 1 \end{pmatrix} \right| = 1,$$

we have

$$|\Theta^*| = \left| \begin{pmatrix} \Theta^{(d)} & \mathbf{0} \\ \mathbf{0}^T & \frac{1}{M} \end{pmatrix} \right| = \frac{1}{M} |\Theta^{(d)}|.$$

□

*Proof of Proposition 3.3.* The aggregated negative log-likelihood function can be written as

$$\begin{aligned} & \frac{1}{d} \sum_{k=1}^d \left\{ -\log |\Theta^{(k)}| + \text{tr} \left( S_k \Theta^{(k)} \right) \right\} \\ &= \frac{1}{d} \sum_{k=1}^d \left\{ -\log |\Theta^{(k)}| + \text{tr} \left( \tilde{S}_k \Theta \right) \right\} \\ &= \frac{1}{d} \sum_{k=1}^d \left\{ -\log |\Theta^{(k)}| \right\} + \text{tr} \left( \left( \frac{1}{d} \sum_{k=1}^d \tilde{S}_k \right) \Theta \right) \\ &= -\log |\Theta^*| + \log(M) + \text{tr} \left( \left( \frac{1}{d} \sum_{k=1}^d \tilde{S}_k - \left( \frac{1}{d^3} \sum_{k=1}^d \mathbf{1}^T \tilde{S}_k \mathbf{1} \right) \mathbf{1} \mathbf{1}^T + M \cdot \mathbf{1} \mathbf{1}^T \right) \Theta \right) \\ &= -\log |\Theta^*| + \log(M) + \text{tr} \left( S^* \left( \Theta^* - \frac{1}{d^2 M} \mathbf{1} \mathbf{1}^T \right) \right). \end{aligned}$$

□

## B Proof of Proposition 4.1

*Proof of Proposition 4.1.* We intend to apply Theorem 3 in Engelke et al. (2022). For that purpose, we first verify all assumptions needed for that theorem, namely Assumptions 1 and 2 therein.

We handle Assumption 1 first. Based on Lemma S3 in Engelke et al. (2022), Condition 4.2 implies that for any  $\xi'' > 0$ , there exists  $K_{\xi''} > 0$  depending on  $\underline{\lambda}$  and  $\bar{\lambda}$ , but independent of  $d$ , such that Assumption 4 therein holds. Denote  $K = K' + 2K_{\xi''}$  and  $\xi = \xi' \xi'' / (1 + \xi' + \xi'')$ . Together with Condition 4.1, we get that Assumption 1 in Engelke et al. (2022) holds. In particular, one can choose  $\xi''$  sufficiently large such that  $\xi$  can be any constant satisfying  $\xi < \xi'$ .

Next, Assumption 2 in Engelke et al. (2022) holds automatically for all non-degenerate HR distribution satisfying our Condition 4.2. Therefore, we can then apply Theorem 3 therein to obtain that there exists positive constants  $C_1$ ,  $C_2$  and  $C_3$ , independent of  $d$ , such that for any  $k_n \geq n^\xi$  and  $\lambda \leq \sqrt{k_n} / (\log n)^4$ ,

$$\mathbb{P} \left( \max_{1 \leq k \leq d} \|\hat{\Sigma}^{(k)} - \Sigma^{(k)}\|_\infty > C_1 \left\{ \left( \frac{k_n}{n} \right)^\xi \left( \log \left( \frac{k_n}{n} \right) \right)^2 + \frac{1 + \lambda}{\sqrt{k_n}} \right\} \right) \leq C_2 d^3 e^{-C_3 \lambda^2}. \quad (\text{B.1})$$

Notice that the constant  $C_1$  here equals to  $\frac{3}{2} \bar{C}$  in Theorem 3 in Engelke et al. (2022) because we are estimating the matrix  $\Sigma$  instead of the variogram  $\Gamma$ .

For any  $\varepsilon \geq C_2 d^3 \exp\{-\frac{C_3 k_n}{(\log n)^8}\}$ , one can choose  $\lambda = \sqrt{\frac{1}{C_3} \log(C_2 d^3 / \varepsilon)} \leq \frac{\sqrt{k_n}}{(\log n)^4}$  in (B.1) to obtain the element-wise bound for the estimation error  $\hat{\Sigma}^{(k)} - \Sigma^{(k)}$  uniformly for all  $1 \leq k \leq d$ .

Since

$$S - \Sigma = \frac{1}{d} \sum_{k=1}^d (\hat{\Sigma}^{(k)} - \Sigma^{(k)}) - \frac{1}{d} \sum_{k=1}^d \left( \frac{1}{d^2} \mathbf{1}^T (\hat{\Sigma}^{(k)} - \Sigma^{(k)}) \mathbf{1} \right) \mathbf{1} \mathbf{1}^T,$$

which implies that  $\|S - \Sigma\|_\infty \leq 2 \max_{1 \leq k \leq d} \|\hat{\Sigma}^{(k)} - \Sigma^{(k)}\|_\infty$ , we immediately get the inequality (4.1) with replacing  $C_1$  by  $2C_1$ . W.l.o.g., we continue using  $C_1$ .

For the asymptotic statement, note that if  $(\log n)^4 \sqrt{\frac{\log d}{k_n}} \rightarrow 0$  as  $n \rightarrow \infty$ , then the lower bound for  $\varepsilon$ ,  $\varepsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ . The asymptotic statement follows immediately.  $\square$

## C Proof of Theorem 4.2

*Proof of Theorem 4.2.* We shall work with the event

$$A_\varepsilon = \{\|S - \Sigma\|_\infty < \delta_n\},$$

which satisfies  $\mathbb{P}(A_\varepsilon) > 1 - \varepsilon$  following Proposition 4.1. Denote  $c := \frac{1}{d^2 M}$ . Then

$$\Theta^* = \Theta + c \mathbf{1} \mathbf{1}^T.$$

Here we omit  $M$  in the notation for simplicity.

Recall that  $\hat{\Theta}^*$  is the solution to the following graphical lasso problem

$$\hat{\Theta}^* := \arg \min_{\Theta^*} \left\{ -\log |\Theta^*| + \text{tr}(S^* \Theta^*) + \gamma_n \sum_{i \neq j} |\Theta_{ij}^* - c| \right\}$$

and  $\hat{\Theta}_{lasso} := \hat{\Theta}^* - c \mathbf{1} \mathbf{1}^T$ . The estimated edge set is  $\hat{E} := \{(i, j) : \hat{\Theta}_{lasso, ij} \neq 0\}$ .

We aim to prove that on  $A_\varepsilon$ ,  $\hat{E} \subset E$ , and

$$\|\hat{\Theta}_{lasso} - \Theta\|_\infty \leq r_n,$$

which is equivalent to proving

$$\|\hat{\Theta}^* - \Theta^*\|_\infty \leq r_n.$$

We first show that the solution  $\hat{\Theta}^*$  exists and is unique. The proof follows the same lines as that for Lemma 3 in Ravikumar et al. (2011). Note that the estimator  $S^*$  is positive definite with all diagonal elements being positive. The rest of the proof follows exactly the same arguments therein.

Next, the solution  $\hat{\Theta}^*$  must satisfy the following KKT condition.

$$-\left(\hat{\Theta}^*\right)^{-1} + S^* + \gamma_n \hat{Z} = 0, \tag{C.1}$$

where

$$\hat{Z}_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \text{sign}(\hat{\Theta}_{ij}^* - c) & \text{if } i \neq j \text{ and } \hat{\Theta}_{ij}^* \neq c, \\ \in [-1, 1] & \text{if } i \neq j \text{ and } \hat{\Theta}_{ij}^* = c. \end{cases}$$

We shall construct a “witness” precision matrix  $\tilde{\Theta}^*$  as follows. Let  $\tilde{\Theta}^*$  be the solution to the following optimization problem,

$$\tilde{\Theta}^* := \arg \min_{\{\Theta^*: \Theta_{ij}^* = c, (i,j) \in E^c\}} -\log |\Theta^*| + \text{tr}(S^* \Theta^*) + \gamma_n \sum_{i \neq j} |\Theta_{ij}^* - c|. \quad (\text{C.2})$$

This is the same optimization but constrained to a smaller domain. Let  $\tilde{E}$  denote the graph recovered from  $\tilde{\Theta}^*$ . Clearly,  $\tilde{\Theta}^*$  satisfies:  $\tilde{\Theta}_{ij}^* = c$  for  $(i, j) \in E^c$ , i.e.  $\tilde{E} \subset E$ .

We shall show that under the conditions in Theorem 4.2,

- $\tilde{\Theta}^*$  satisfies the above KKT condition;
- $\|\tilde{\Theta}^* - \Theta^*\|_\infty \leq r_n$ .

Then by uniqueness,  $\tilde{\Theta}^* = \hat{\Theta}^*$  and satisfies the goal that we are aiming to prove.

With a similar argument regarding the existence and uniqueness of the original optimization problem, the solution to the problem (C.2) also exists and is unique. In addition, it satisfies a similar KKT condition as follows,

$$-\left(\tilde{\Theta}^*\right)_{ij}^{-1} + S_{ij}^* + \gamma_n \tilde{Z}_{ij} = 0, \quad (i, j) \in E,$$

where

$$\tilde{Z}_{ij} = \begin{cases} \text{sign}(\tilde{\Theta}_{ij}^* - c) & \text{if } \tilde{\Theta}_{ij}^* \neq c, \\ \in [-1, 1] & \text{if } \tilde{\Theta}_{ij}^* = c. \end{cases}$$

Note that this coincides with the KKT condition (C.1), but only on entries indexed by  $E$ . As a matter of fact,  $\tilde{Z}$  is only defined on  $E$ . In order to argue  $\tilde{\Theta}^*$  as a candidate for  $\hat{\Theta}^*$  and satisfies the full KKT condition, we will now extend the definition of  $\tilde{Z}$  to  $E^c$  as well.

Define

$$\tilde{Z}_{ij} := \frac{1}{\gamma_n} \left( \left(\tilde{\Theta}^*\right)_{ij}^{-1} - S_{ij}^* \right), \quad (i, j) \notin E.$$

Then the pair  $(\tilde{\Theta}^*, \tilde{Z})$  satisfies the original KKT equation (C.1). What remains to be proved is that  $\tilde{Z}$  also satisfies

$$|\tilde{Z}_{ij}| \leq 1, \quad (i, j) \notin E.$$

To summarize, in order to complete the proof of Theorem 4.2, we will show that on the set  $A_e$ ,

**Goal 1:**  $|\tilde{Z}_{ij}| \leq 1, \quad (i, j) \notin E$ .

**Goal 2:**  $\|\tilde{\Theta}^* - \Theta^*\|_\infty \leq r_n$ .

In the rest of the proof we denote

$$\Delta := \tilde{\Theta}^* - \Theta^*.$$

Note that for  $(i, j) \notin E$ ,  $\tilde{\Theta}_{ij}^* = c$  by definition and  $\Theta_{ij} = c$ . Therefore,  $\Delta_{E^c} = \mathbf{0}$  and **Goal 2** above can be translated to

$$\|\Delta_E\|_\infty \leq r_n.$$

To handle the KKT condition for  $\tilde{\Theta}^*$ , we start with handling  $(\tilde{\Theta}^*)^{-1}$  as follows:

$$\begin{aligned} (\tilde{\Theta}^*)^{-1} &= (\Theta^* + \Delta)^{-1} \\ &= (\Theta^*(I + \Sigma^*\Delta))^{-1} \\ &= (I + \Sigma^*\Delta)^{-1}\Sigma^* =: J\Sigma^*. \end{aligned}$$

Now in the case where  $\|\Sigma^*\Delta\|_\infty < 1$ , we can expand  $J$  as

$$J = \sum_{k=0}^{\infty} (-1)^k (\Sigma^*\Delta)^k = I - \Sigma^*\Delta + (\Sigma^*\Delta)^2 J.$$

Inspired from this relation, we can use  $\Sigma^* - \Sigma^*\Delta\Sigma^*$  to approximate  $(\tilde{\Theta}^*)^{-1}$  and define

$$R := (\tilde{\Theta}^*)^{-1} - (\Sigma^* - \Sigma^*\Delta\Sigma^*),$$

as the approximation error. Note that  $R$  is defined regardless of whether  $\|\Sigma^*\Delta\|_\infty < 1$ .

Recall that  $\Sigma^* = \Sigma + M\mathbf{1}\mathbf{1}^T$  and  $S^* = S + M\mathbf{1}\mathbf{1}^T$ . Define

$$R' := S^* - \Sigma^* = S - \Sigma.$$

On the set  $A_\varepsilon$ , we have that  $\|R'\|_\infty \leq \delta_n$ .

Rewrite the KKT condition as

$$\Sigma^*\Delta\Sigma^* - R + R' + \gamma_n \tilde{Z} = 0.$$

We vectorize it using the notation  $\bar{\cdot}$  as the vectorization of a matrix. Then the vectorized KKT condition is

$$\overline{\Sigma^*\Delta\Sigma^*} - \bar{R} + \bar{R}' + \gamma_n \bar{\tilde{Z}} = 0.$$

Note that

$$\overline{\Sigma^*\Delta\Sigma^*} = (\Sigma^* \otimes \Sigma^*)\bar{\Delta} =: \Omega\bar{\Delta},$$

where  $\Omega := \Sigma^* \otimes \Sigma^*$  denotes the Kronecker product of  $\Sigma^*$  with itself. Then we have

$$\Omega\bar{\Delta} - \bar{R} + \bar{R}' + \gamma_n \bar{\tilde{Z}} = 0.$$

By examining the rows of  $\Omega$  indexed by  $E$  and  $E^c$  separately and noting that  $\Delta_{E^c} = 0$ , we get

$$\Omega_{EE}\bar{\Delta}_E - \bar{R}_E + \bar{R}'_E + \gamma_n \bar{\tilde{Z}}_E = 0, \quad (\text{C.3})$$

$$\Omega_{E^cE}\bar{\Delta}_E - \bar{R}_{E^c} + \bar{R}'_{E^c} + \gamma_n \bar{\tilde{Z}}_{E^c} = 0. \quad (\text{C.4})$$

## Proof of Goal 2

To prove Goal 2, we shall show that for any pre-specified  $\|\tilde{Z}_E\|_\infty \leq 1$ , there exists a solution to  $\Delta$  which satisfies:

- $-(\Theta^* + \Delta)_E^{-1} + S_E^* + \gamma_n \tilde{Z}_E = 0$ ;
- $\Delta_{E^c} = \mathbf{0}$ ;
- $\|\Delta_E\|_\infty \leq r_n$ .

With the statement above proven, given the  $\tilde{Z}_E$  produced from the KKT condition for  $\tilde{\Theta}^*$ , a solution  $\Delta$  exists and coincides with  $\tilde{Z}_E$ . Then this is the unique solution. Hence  $\|\tilde{\Theta}^* - \Theta^*\|_\infty \leq r_n$  which concludes **Goal 2**.

Now we construct such a solution  $\Delta$ . Recall that  $\Delta_{E^c} = \mathbf{0}$ , we only need to construct a suitable  $\Delta_E$  by utilizing the Brouwer fixed point theorem.

The solution  $\Delta_E$  satisfies (C.3), which can be rewritten as

$$\bar{\Delta}_E = (\Omega_{EE})^{-1} \left( \bar{R}_E - \bar{R}'_E - \gamma_n \tilde{Z}_E \right).$$

We regard

$$R = \left( \tilde{\Theta}^* \right)^{-1} - (\Sigma^* - \Sigma^* \Delta \Sigma^*)$$

as a function of  $\Delta$  or eventually a function of  $\bar{\Delta}_E$ . To stress this point we define it as  $R = R(\Delta_E)$ . Also recall that  $R' = S^* - \Sigma^*$  does not depend on  $\bar{\Delta}_E$ . Then we can write the above equation as

$$\bar{\Delta}_E = (\Omega_{EE})^{-1} \left( \bar{R}_E(\bar{\Delta}_E) - \bar{R}'_E - \gamma_n \tilde{Z}_E \right) := F(\bar{\Delta}_E).$$

Consider the closed ball  $B(r_n) := \{x \in \mathbb{R}^{|E|} : \|x\|_\infty \leq r_n\}$ . If  $F$  is a continuous mapping from  $B(r_n)$  onto itself, then there exists a fixed point  $\bar{\Delta}_E$  on  $B(r_n)$  such that  $\bar{\Delta}_E = F(\bar{\Delta}_E)$  following the Brouwer fixed point theorem. This is exactly the desired solution. Since  $F$  is clearly continuous, we only need to show that  $F$  projects  $B(r_n)$  onto itself, that is, for any  $\bar{\Delta}_E$  satisfying  $\|\bar{\Delta}_E\|_\infty \leq r_n$ , we have  $\|F(\bar{\Delta}_E)\|_\infty \leq r_n$ .

Assume that  $\|\bar{\Delta}_E\|_\infty \leq r_n$ . We write

$$\|F(\bar{\Delta}_E)\|_\infty \leq \|(\Omega_{EE})^{-1}\|_\infty \left( \|R\|_\infty + \|R'\|_\infty + \gamma_n \|\tilde{Z}_E\|_\infty \right) \leq \|(\Omega_{EE})^{-1}\|_\infty (\|R\|_\infty + \|R'\|_\infty + \gamma_n),$$

due to the fact that  $\|\tilde{Z}_E\|_\infty \leq 1$

We first handle  $\|R\|_\infty$ . Recall that

$$R := \left( \tilde{\Theta}^* \right)^{-1} - (\Sigma^* - \Sigma^* \Delta \Sigma^*) = \sum_{k=2}^{\infty} (-1)^k (\Sigma^* \Delta)^k \Sigma^* = (\Sigma^* \Delta)^2 J \Sigma^*,$$

where

$$J = (I + \Sigma^* \Delta)^{-1} = \sum_{k=0}^{\infty} (-1)^k (\Sigma^* \Delta)^k,$$

provided that  $|||\Sigma^* \Delta|||_\infty < 1$ . To ensure this condition, note that

$$|||\Sigma^* \Delta|||_\infty = |||\Sigma^*(\tilde{\Theta} - \Theta)|||_\infty \leq D |||\Sigma^*|||_\infty \cdot r_n,$$

where  $D$  is the maximum degree in the graph. Therefore  $|||\Sigma^* \Delta|||_\infty < 1$  holds by requiring that

$$D |||\Sigma^*|||_\infty \cdot r_n \leq C_4 < 1. \quad (\text{C.5})$$

The upper bound  $C_4$  implies that  $|||J^T|||_\infty \leq \frac{1}{1-C_4}$ .

With the condition (C.5), we can further derive an upper bound for  $\|R\|_\infty$ . Consider one specific element in  $R$ . With denoting  $\mathbf{e}_i$  as a vector with all zero elements except a one element at the  $i$ -th dimension, we have that

$$R_{ij} = \mathbf{e}_i^T (\Sigma^* \Delta)^2 J \Sigma^* \mathbf{e}_j \leq \|\mathbf{e}_i^T (\Sigma^* \Delta)^2\|_\infty \|J \Sigma^* \mathbf{e}_j\|_1 \leq \|(\Sigma^* \Delta)^2\|_\infty |||\Sigma^* J^T|||_\infty.$$

By considering all possible  $(i, j)$  we get that,

$$\begin{aligned} \|R\|_\infty &\leq \|(\Sigma^* \Delta)^2\|_\infty |||\Sigma^* J^T|||_\infty \\ &\leq |||\Sigma^* \Delta|||_\infty \|\Sigma^* \Delta\|_\infty |||J^T|||_\infty |||\Sigma^*|||_\infty \\ &\leq D |||\Sigma^*|||_\infty \cdot r_n \cdot |||\Sigma^*|||_\infty r_n \cdot \frac{1}{1-C_4} \cdot |||\Sigma^*|||_\infty \\ &< C_5 \cdot r_n^2, \end{aligned}$$

where  $C_5 = \frac{D}{1-C_4} |||\Sigma^*|||_\infty^3$ .

Next, since  $R' = S^* - \Sigma^* = S - \Sigma$ , we have that on  $A_\varepsilon$ ,  $\|R'\|_\infty \leq \delta_n$ . Combining the upper bounds for  $\|R\|_\infty$  and  $\|R'\|_\infty$ , we get that on  $A_\varepsilon$ ,

$$\|F(\Delta_E)\|_\infty \leq |||(\Omega_{EE})^{-1}|||_\infty (C_5 \cdot r_n^2 + \delta_n + \gamma_n) \leq r_n,$$

by requiring that

$$C_5 \cdot r_n^2 + \delta_n + \gamma_n \leq \frac{1}{|||(\Omega_{EE})^{-1}|||_\infty} \cdot r_n \quad (\text{C.6})$$

If the two required conditions (C.5) and (C.6) hold, we achieve **Goal 2** by utilizing the Brouwer fixed point theorem.

## Proof of Goal 1

To prove **Goal 1**, we shall show that with the constructed solution above, we have

$$\|\bar{Z}_{E^c}\|_\infty \leq 1.$$

We rewrite the equation (C.4) as

$$\bar{Z}_{E^c} = -\frac{1}{\gamma_n} \Omega_{EE^c} \bar{\Delta}_E + \frac{1}{\gamma_n} \bar{R}_{E^c} - \frac{1}{\gamma_n} \bar{R}'_{E^c},$$

and the substitute  $\bar{\Delta}_E$  above using (C.3) to get that

$$\bar{\bar{Z}}_{E^c} = -\frac{1}{\gamma_n} \Omega_{EE^c} (\Omega_{EE})^{-1} \left( -\bar{R}_E + \bar{R}'_E + \gamma_n \bar{\bar{Z}}_E \right) + \frac{1}{\gamma_n} \bar{R}_{E^c} - \frac{1}{\gamma_n} \bar{R}'_{E^c}.$$

The upper bound for  $\|\bar{\bar{Z}}_{E^c}\|_\infty$  is then

$$\|\bar{\bar{Z}}_{E^c}\|_\infty \leq \frac{1}{\gamma_n} \|\Omega_{EE^c} (\Omega_{EE})^{-1}\|_\infty \left( \|R\|_\infty + \|R'\|_\infty + \gamma_n \|\bar{\bar{Z}}_E\|_\infty \right) + \frac{1}{\gamma_n} \|R\|_\infty + \frac{1}{\gamma_n} \|R'\|_\infty.$$

Using the same upper bounds derived in the proof of **Goal 2**, we have

$$\begin{aligned} \|\bar{\bar{Z}}_{E^c}\|_\infty &\leq \frac{1}{\gamma_n} \|\Omega_{EE^c} (\Omega_{EE})^{-1}\|_\infty (\delta_n + C_5 \cdot r_n^2 + \gamma_n) + \frac{1}{\gamma_n} (\delta_n + C_5 \cdot r_n^2) \\ &= \|\Omega_{EE^c} (\Omega_{EE})^{-1}\|_\infty + \frac{1}{\gamma_n} (\|\Omega_{EE^c} (\Omega_{EE})^{-1}\|_\infty + 1) (\delta_n + C_5 \cdot r_n^2). \end{aligned}$$

Since Condition (4.3) ensures that  $\|\Omega_{EE^c} (\Omega_{EE})^{-1}\|_\infty < 1 - \alpha$ , To satisfy  $\|\bar{\bar{Z}}_{E^c}\|_\infty \leq 1$ , we only need to further require

$$\delta_n + C_5 \cdot r_n^2 \leq \frac{\alpha}{1 - \alpha} \gamma_n. \quad (\text{C.7})$$

To conclude, the theorem is proven provided that the three conditions (C.5)–(C.7) hold. The last step is to verify these three relations under the conditions in Theorem 4.2.

Recall that  $r_n$  is defined in (4.4)

$$r_n := \frac{\|\Omega_{EE}^{-1}\|_\infty}{1 - \alpha} \cdot \gamma_n.$$

Clearly, this definition together with (C.7) implies (C.6). Hence we only need to verify the conditions (C.5) and (C.7).

We write the two conditions in terms of  $\delta_n$  and  $\gamma_n$ :

$$\begin{aligned} D \|\Sigma^*\|_\infty \frac{\|\Omega_{EE}^{-1}\|_\infty}{1 - \alpha} \cdot \gamma_n &\leq C_4 \\ \delta_n + \frac{D}{1 - C_4} \|\Sigma^*\|_\infty^3 \cdot \left( \frac{\|\Omega_{EE}^{-1}\|_\infty}{1 - \alpha} \right)^2 \cdot \gamma_n^2 &\leq \frac{\alpha}{1 - \alpha} \cdot \gamma_n. \end{aligned}$$

where  $C_5$  is substituted by  $\frac{D}{1 - C_4} \|\Sigma^*\|_\infty^3$ :

Note that the lower bound for  $\gamma_n$  in (4.3) ensures that  $\delta_n \leq \epsilon \frac{\alpha}{1 - \alpha} \gamma_n$  for some  $0 < \epsilon < 1$ , we thus need to require that

$$\frac{D}{1 - C_4} \|\Sigma^*\|_\infty^3 \cdot \left( \frac{\|\Omega_{EE}^{-1}\|_\infty}{1 - \alpha} \right)^2 \cdot \gamma_n \leq (1 - \epsilon) \frac{\alpha}{1 - \alpha},$$

which guarantees the second condition. Together with the first condition, we have obtained an upper bound for  $\gamma_n$  as

$$\gamma_n \leq \min \left\{ \frac{C_4(1 - \alpha)}{D \|\Sigma^*\|_\infty \|\Omega_{EE}^{-1}\|_\infty}, \frac{(1 - C_4)(1 - \epsilon)\alpha(1 - \alpha)}{D \|\Sigma^*\|_\infty^3 \|\Omega_{EE}^{-1}\|_\infty^2} \right\}.$$

We choose  $C_4$  such that the two terms in the minimum are equal. That is

$$C_4 = \frac{(1 - \epsilon)\alpha}{(1 - \epsilon)\alpha + \|\Sigma^*\|_\infty^2 \|(\Omega_{EE})^{-1}\|_\infty} < 1,$$

which leads to

$$\gamma_n \leq \frac{(1 - \epsilon)\alpha(1 - \alpha)}{D \|\Sigma^*\|_\infty \|(\Omega_{EE})^{-1}\|_\infty [(1 - \epsilon)\alpha + \|\Sigma^*\|_\infty^2 \|(\Omega_{EE})^{-1}\|_\infty]}.$$

This is exactly the required upper bound for  $\gamma_n$  in (4.2). □

## D Blockwise coordinate descent algorithm

For the sake of clarify, we abuse the notations in this section by using  $\Theta$  for  $\Theta^*$  and  $S$  for  $S^*$ . We describe the algorithm to solve the minimization problem:

$$\min_{\Theta \geq 0} -\log |\Theta| + \text{tr}(S\Theta) + \gamma \sum_{i \neq j} |\Theta_{ij} - c|.$$

Here  $\Theta$  is the precision matrix to be estimated and  $S$  is an estimated covariance matrix guaranteed to be positive definite.

Similar to classical graphical lasso, the objective function is convex. Searching for the optimum is equivalent to solving the KKT condition

$$-\Theta^{-1} + S + \gamma \mathbf{Z} = \mathbf{0},$$

where  $\mathbf{Z}$  is a matrix of component-wise signs of  $\Theta - c\mathbf{1}\mathbf{1}^\top$ :

$$\begin{aligned} z_{ii} &= 0 & \text{if } i &= j \\ z_{ij} &= \text{sign}(\theta_{ij} - c) & \text{if } i &\neq j, \quad \theta_{ij} \neq c \\ z_{ij} &\in [-1, 1] & \text{if } i &\neq j, \quad \theta_{ij} = c. \end{aligned}$$

In the following we will demonstrate a blockwise coordinate descent approach to solve this problem. A primitive version is used for the conventional graphical lasso problem for  $c = 0$  in Friedman et al. (2008) and implemented in the R package *glasso*. However, to apply that algorithm to our generalized problem, an additional matrix inversion of a  $(d - 1) \times (d - 1)$  matrix is required at each iteration step. By contrast, the algorithm in this appendix can be seen as the dual problem of that in Friedman et al. (2008). Similar to Mazumder and Hastie (2012), no matrix inversion is needed in our algorithm.

### D.1 The idea

Let us consider solving for  $\Sigma = \Theta^{-1}$ . Then the KKT condition becomes

$$-\Sigma + S + \gamma \mathbf{Z} = \mathbf{0}. \tag{D.1}$$

Since  $z_{ii} = 0$  for each  $i$ , we first get that

$$\hat{\sigma}_{ii} = s_{ii}, \quad \forall i = 1, \dots, d.$$

Let us write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \boldsymbol{\sigma}_{1d} \\ \boldsymbol{\sigma}_{1d}^T & \sigma_{dd} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \Theta_{11} & \boldsymbol{\theta}_{1d} \\ \boldsymbol{\theta}_{1d}^T & \theta_{dd} \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{z}_{1d} \\ \mathbf{z}_{1d}^T & z_{dd} \end{pmatrix}.$$

In the following, we aim to update  $\Sigma$  by keeping  $\Theta_{11}$  fixed. We iterate through the columns (rows) of  $\Theta$  until a convergence is reached.

From  $\Sigma \cdot \Theta = I$ , we have the following two presentations of  $\Sigma$ :

$$\begin{pmatrix} \Sigma_{11} & \boldsymbol{\sigma}_{1d} \\ \boldsymbol{\sigma}_{1d}^T & \sigma_{dd} \end{pmatrix} = \begin{pmatrix} (\Theta_{11} - \boldsymbol{\theta}_{1d}\boldsymbol{\theta}_{1d}^T)^{-1} & -\theta_{dd}^{-1}\Sigma_{11}\boldsymbol{\theta}_{1d} \\ \cdot & \theta_{dd}^{-1} - \theta_{dd}^{-2}\boldsymbol{\theta}_{1d}^T\Sigma_{11}\boldsymbol{\theta}_{1d} \end{pmatrix} \quad (\text{D.2})$$

$$= \begin{pmatrix} \Theta_{11}^{-1} + \frac{\Theta_{11}^{-1}\boldsymbol{\theta}_{1d}\boldsymbol{\theta}_{1d}^T\Theta_{11}^{-1}}{\theta_{dd} - \boldsymbol{\theta}_{1d}^T\Theta_{11}^{-1}\boldsymbol{\theta}_{1d}} & -\frac{\Theta_{11}^{-1}\boldsymbol{\theta}_{1d}}{\theta_{dd} - \boldsymbol{\theta}_{1d}^T\Theta_{11}^{-1}\boldsymbol{\theta}_{1d}} \\ \cdot & \frac{1}{\theta_{dd} - \boldsymbol{\theta}_{1d}^T\Theta_{11}^{-1}\boldsymbol{\theta}_{1d}} \end{pmatrix} \quad (\text{D.3})$$

where  $\cdot$  denotes the mirroring of elements in the upper triangle. The proofs can be found in Section D.2. The same formula can be applied for a representation of  $\Theta$  using  $\Sigma$ .

Consider the last column of (D.1), we get

$$-\boldsymbol{\sigma}_{1d} + \mathbf{s}_{1d} + \gamma\mathbf{z}_{1d} = \mathbf{0}.$$

Plugging in (D.3), we have

$$\sigma_{dd}\Theta_{11}^{-1}\boldsymbol{\theta}_{1d} + \mathbf{s}_{1d} + \gamma\mathbf{z}_{1d} = \mathbf{0},$$

where  $\sigma_{dd}$  is known. Now set  $\boldsymbol{\beta} = (\boldsymbol{\theta}_{1d} - c\mathbf{1})\sigma_{dd}$ . Then the above equation becomes

$$\Theta_{11}^{-1}(\boldsymbol{\beta} + c\sigma_{dd} \cdot \mathbf{1}) + \mathbf{s}_{1d} + \gamma\mathbf{z}_{1d} = \Theta_{11}^{-1}\boldsymbol{\beta} + c\sigma_{dd} \cdot \Theta_{11}^{-1}\mathbf{1} + \mathbf{s}_{1d} + \gamma\mathbf{z}_{1d} = \mathbf{0}, \quad (\text{D.4})$$

where we aim to solve for  $\boldsymbol{\beta}$ . Note that

$$\mathbf{z}_{1d} = \text{sign}(\boldsymbol{\theta}_{1d} - c\mathbf{1}) = \text{sign}((\boldsymbol{\theta}_{1d} - c\mathbf{1})\sigma_{dd}) = \text{sign}(\boldsymbol{\beta}).$$

Then solving for (D.4) is equivalent to the standard quadratic lasso problem:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}; \Theta_{11}, \sigma_{dd}, \mathbf{s}_{1d}, \gamma) = \frac{1}{2}\boldsymbol{\beta}^T\Theta_{11}^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^T(c\sigma_{dd} \cdot \Theta_{11}^{-1}\mathbf{1} + \mathbf{s}_{1d}) + \gamma\|\boldsymbol{\beta}\|_1, \quad (\text{D.5})$$

which can be solve efficiently using elementwise coordinate descent if we know  $\Theta_{11}^{-1}$ .

At each iteration, we aim to update  $\Theta$  and then  $\Sigma$ . Given  $\Theta$  and  $\Sigma$  from the previous iteration, we proceed as follows:

- Calculate  $\Theta_{11}^{-1}$  from

$$\Theta_{11}^{-1} = \Sigma_{11} - \sigma_{dd}^{-1}\boldsymbol{\sigma}_{1d}\boldsymbol{\sigma}_{1d}^T.$$

This is the opposite representation of (D.3).

- Update  $\boldsymbol{\theta}_{1d}$ : Solve for  $\hat{\boldsymbol{\beta}}$  from (D.5) and update

$$\boldsymbol{\theta}_{1d} \leftarrow \sigma_{dd}^{-1} \hat{\boldsymbol{\beta}} + c\mathbf{1}.$$

- Update  $\theta_{dd}$ :

$$\theta_{dd} = \sigma_{dd}^{-1} + \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \boldsymbol{\theta}_{1d}.$$

This comes from the representation in (D.3).

- Update the entire  $\Sigma$  matrix from representation (D.3) using the fixed  $\Theta_{11}^{-1}$  and the updated  $\boldsymbol{\theta}_{1d}$  and  $\theta_{dd}$ .

The algorithm is summarized in Algorithm 1.

---

**Algorithm 1** Graphical lasso algorithm for extremes

---

Input  $c$ ,  $S$  and  $\gamma$ .

1. Initialize  $\Sigma = S$  and  $\Theta = S^{-1}$ .
2. In each iteration, update  $\Sigma$  while keeping a  $(d-1) \times (d-1)$  submatrix of  $\Theta$  fixed. Iterate through the columns repeatedly on the following steps until convergence.

(a) Rearrange the rows/columns such that the target column is last (implicitly).

(b) Calculate

$$\Theta_{11}^{-1} = \Sigma_{11} - \sigma_{dd}^{-1} \boldsymbol{\sigma}_{1d} \boldsymbol{\sigma}_{1d}^T.$$

(c) Solve for (D.5)

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}; \Theta_{11}, \sigma_{dd}, \mathbf{s}_{1d}, \gamma)$$

(d) Update  $\boldsymbol{\theta}_{1d}$ :

$$\boldsymbol{\theta}_{1d} \leftarrow \sigma_{dd}^{-1} \hat{\boldsymbol{\beta}} + c\mathbf{1}.$$

(e) Update  $\theta_{dd}$ :

$$\theta_{dd} = \sigma_{dd}^{-1} + \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \boldsymbol{\theta}_{1d}.$$

(f) Update the entire  $\Sigma$  matrix from representation (D.3) using the fixed  $\Theta_{11}^{-1}$  and the updated  $\boldsymbol{\theta}_{1d}$  and  $\theta_{dd}$ .

---

## D.2 Proofs

We will now prove the equations (D.2) and (D.3).

From  $\Sigma \cdot \Theta = I$ , we get the following equations.

$$\Sigma_{11} \Theta_{11} + \boldsymbol{\sigma}_{1d} \boldsymbol{\theta}_{1d}^T = I \tag{D.6}$$

$$\Sigma_{11} \boldsymbol{\theta}_{1d} + \theta_{dd} \cdot \boldsymbol{\sigma}_{1d} = \mathbf{0} \tag{D.7}$$

$$\boldsymbol{\sigma}_{1d}^T \Theta_{11} + \sigma_{dd} \cdot \boldsymbol{\theta}_{1d}^T = \mathbf{0}^T \tag{D.8}$$

$$\boldsymbol{\sigma}_{1d}^T \boldsymbol{\theta}_{1d} + \sigma_{dd} \cdot \theta_{dd} = 1. \quad (\text{D.9})$$

From (D.7), we have

$$\boldsymbol{\sigma}_{1d} = -\theta_{dd}^{-1} \Sigma_{11} \boldsymbol{\theta}_{1d}.$$

From (D.9), we have

$$\sigma_{dd} = \theta_{dd}^{-1} (1 - \boldsymbol{\sigma}_{1d}^T \boldsymbol{\theta}_{1d}) = \theta_{dd}^{-1} + \theta_{dd}^{-2} \boldsymbol{\theta}_{1d}^T \Sigma_{11} \boldsymbol{\theta}_{1d}.$$

From (D.6), we have

$$\Sigma_{11}^{-1} = \Theta_{11} + \Sigma_{11}^{-1} \boldsymbol{\sigma}_{1d} \boldsymbol{\theta}_{1d}^T = \Theta_{11} - \Sigma_{11}^{-1} \theta_{dd}^{-1} \Sigma_{11} \boldsymbol{\theta}_{1d} \boldsymbol{\theta}_{1d}^T = \Theta_{11} - \theta_{dd}^{-1} \boldsymbol{\theta}_{1d} \boldsymbol{\theta}_{1d}^T$$

This proves (D.2).

Now consider (D.3). From (D.6) and (D.8), we have

$$\Sigma_{11} = \Theta_{11}^{-1} - \boldsymbol{\sigma}_{1d} \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \quad (\text{D.10})$$

$$\boldsymbol{\sigma}_{1d} = -\sigma_{dd} \Theta_{11}^{-1} \boldsymbol{\theta}_{1d} \quad (\text{D.11})$$

Plug (D.11) into (D.9), we get

$$\sigma_{dd} \theta_{dd} - \sigma_{dd} \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \boldsymbol{\theta}_{1d} = 1$$

and hence

$$\sigma_{dd} = \frac{1}{\theta_{dd} - \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \boldsymbol{\theta}_{1d}}.$$

Plugging in (D.11), we have

$$\boldsymbol{\sigma}_{1d} = -\frac{\Theta_{11}^{-1} \boldsymbol{\theta}_{1d}}{\theta_{dd} - \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \boldsymbol{\theta}_{1d}}.$$

Plugging in (D.10), we have

$$\Sigma_{11} = \Theta_{11}^{-1} + \frac{\Theta_{11}^{-1} \boldsymbol{\theta}_{1d} \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1}}{\theta_{dd} - \boldsymbol{\theta}_{1d}^T \Theta_{11}^{-1} \boldsymbol{\theta}_{1d}}.$$

This proves (D.3).