Graphical Learning and Clustering based on Hüsler-Reiss Graphical Models

Peng Zhong, Scott A. Sisson, and Boris Beranger
School of Mathematics and Statistics,
University of New South Wales, Sydney 2052, Australia.
{peng.zhong, scott.sisson, b.beranger}@unsw.edu.au

August 15, 2025

Abstract

Since the introduction of extremal graphical models, various methods have been proposed to learn the underlying dependence structure using graphical lasso-based algorithms. Unlike conditional independence in Gaussian graphical models of dimension d, extremal conditional independence is defined by conditioning on single site being large, which naturally leads to d precision matrices of dimension $d-1$ that encode extremal conditional independence. Recent literature has shown the existence of a single positive semi-definite precision matrix that encodes the extremal conditional independence. However, current inference methods, e.g, EGlearn, relies on learning d sub-graphs and often fail to guarantee connectivity, frequently yielding multiple disconnected components as the extremal graph must be connected and full independence is not permitted.

Additionally, these graphical lasso-based approaches use empirical estimates of either d variogram matrices or a aggregated covariance matrix from d matrices, which can be unintuitive. In this paper, we introduce a novel inference method for the Hüsler-Reiss graphical model based on three different estimators of the precision matrix's pesudo inverse, termed as extremal covariance matrix, and directly exploit the structure of the precision matrix, which mimic the Gaussian case. Comparing to existing methods, we demonstrate that our method is as fast but more accurate for learning the graphical dependence structure, while ensuring connectivity. Furthermore, when full independence is assumed between disconnected components, our approach can also accommodate it, yielding the exactly disconnected components, which provides a new tool for extremal clustering. We also provide a data driven method using hierarchical clustering method to learn the clustered structure. We validate the performance of our method through simulation studies and two real data applications.

Keywords: Graphical extremes; Hüsler-Reiss process; Multivariate Pareto distribution; Graphical lasso; Cluster detection;

1 Introduction

Graphical models are powerful tools for understanding structured relationships in data, particularly in the context of multivariate Gaussian distributions where conditional independence can be directly inferred from the precision matrix (Rue and Held, 2005). However, for multivariate extremes, graphical models have only recently been developed for a class of models known as multivariate Pareto distributions associated with the maximum risk functional (Engelke and Hitz, 2020; Hentschel *et al.*, 2024), with a focus on the Hüsler-Reiss graphical model. Similar to the Gaussian graphical model, it has been shown that the Hüsler-Reiss graphical model has a single precision matrix, termed as extremal precision matrix, that encodes the extremal conditional independence structure, and the corresponding graph should be connected (Hentschel *et al.*, 2024). However, existing inference methods for learning the graphical structure of the Hüsler-Reiss graphical model often fail to guarantee the connectivity and tend to produce multiple disconnected components (Hentschel *et al.*, 2024; Wan and Zhou, 2025; Engelke *et al.*, 2025). In the Gaussian setting, disconnected graphical components are interpreted as full independence, but this interpretation does not hold for the Hüsler-Reiss graphical model, as a single parametrized Hüsler-Reiss model can only capture asymptotic dependence, but not asymptotic independence. Therefore, there is a critical gap between the theoretical understanding and the methods for learning the graphical structure, particularly in dealing with disconnected graphical components, which requires urgent attention.

Current inference methods for graphical extremes with dimension d, including the EGlearn (Engelke *et al.*, 2025; Hentschel *et al.*, 2024) and the extremal graphical lasso (Wan and Zhou, 2025), are based either on empirical variogram matrix, $\widehat{\Gamma}^{(k)}$, or empirical covariance matrix of dimension $d-1$ conditioning on each dimension k, $k = 1, \ldots, d$. The EGlearn method involves using the graphical lasso method with each $\widehat{\Gamma}^{(k)}$ repeatedly and then aggregating the learned d sub-graphs with $d-1$ nodes into a single graph of d nodes, while the extremal graphical lasso method in Wan and Zhou (2025) combines d empirical covariance matrices into a single empirical covariance matrix and then applies the a graphical lasso method to learn a single graph by penalising the entries of the precision matrix towards a positive constant. Besides the lack of connectivity guarantee, both methods are not intuitive, as neither the empirical variogram matrix $\widehat{\Gamma}^{(k)}$ or the empirical covariance matrix $\widehat{\Sigma}^{(k)}$ are directly linked to the graphical structure unlike the precision (covariance) matrix in the Gaussian graphical model. The EGlearn method can be potentially computationally expensive as it requires d times of graphical lasso optimisation to be applied. In contrast, the extremal graphical lasso method only requires one time of graphical lasso optimisation. However, according to the simulation study in Wan and Zhou (2025), the extremal graphical lasso is not as efficient as the EGlearn method and performs poorly in terms of F score, which is a measure for predictive performance of identifying edges in the graphical structure, especially when d is relatively large ($d \geq 100$).

In this paper, we propose a new efficient inference method for learning the graphical structure of the Hüsler-Reiss graphical model while guaranteeing the connectivity, which is directly using 3 different estimators of the extremal precision matrix's pesudo-inverse. We termed this pesudo-inverse of the extremal precision matrix as extremal covariance matrix. We also showed that these extremal covariance matrix estimators, which are based on different risk functionals,are consistent and unbiased, and we further justify our the usage of these estimators by showing that these risk functionals will lead to a Hüsler-Reiss with the same extremal covariance matrix as long as they are in the max-domain attraction of a Hüsler-Reiss max-stable distribution. With the estimated extremal covariance matrix, we then optimise a graphical lasso objective function, assuming the model is extremal multivariate total positive of order 2 (EMTP2) (Röttger *et al.*, 2023), to infer the extremal precision matrix. The assumption of EMTP2 is not restricted at all as one would expect positive associations for extreme random vectors and there are indeed many classic max-stable models preserving the EMTP2 property. For Hüsler-Reiss graphical models, the EMTP2 condition essentially means the extremal precision matrix is a M-matrix, i.e., off-diagonal elements of the extremal precision matrix is non-positive. In the Gaussian setting, assuming the precision matrix is a M-matrix and learning the precision (covariance) matrix has been widely explored in the literature (Lauritzen *et al.*, 2019; Ying *et al.*, 2021; Kumar *et al.*, 2019). However, assuming the extremal precision matrix is a M-matrix will further make the extremal precision matrix a laplacian matrix, i.e., rows of the matrix are summed to zero, which is a subset of the M-matrix family. If the extremal precision matrix is a laplacian matrix, then, we can enforce the connectivity of the learnt graph by imposing spectral constraints on the eigenvalues of the extremal precision matrix. In this paper, we adopt the algorithm from (Kumar *et al.*, 2019) to learn the structured graph via spectral constraints, which only require one time optimisation and is computationally efficient. Other graphical methods with laplacian constraints can also be used potentially, such as Ying *et al.* (2020, 2021), where a generalized lasso penalty is used rather than the L_1 norm penalty on each off-diagonal element of the estimated precision matrix, since simply optimizing the graphical lasso with laplacian constraints will not lead to a sparse graph when the lasso penalty is too large, instead, it will lead to a complete graph (Ying *et al.*, 2020). Röttger *et al.* (2023) used a similar approach as in Ying *et al.* (2021) but without the sparsity regularization to learn a connected graph under EMTP2 assumption. Therefore, their method cannot guarantee sparsity of the learnt graph and it is not designed for structure learning. Moreover, Engelke *et al.* (2025) also showed that the method in Röttger *et al.* (2023) does not perform well, especially in high dimensions.

While independence assumption is made for disconnected components, i.e., the precision matrix is a block diagonal laplacian matrix, our approach can also accommodate it by imposing spectral constraints and learn the exact disconnected components given the correct number of disconnected components, which provides a new tool for clustering independent components. This is not done by direct usage of the estimated extremal covariance matrix since the estimator for the extremal covariance matrix is only designed for a single dependent component. Moreover, since the data we have is usually not already following an extreme distribution, but only assumed in the max-domain attraction of a Hüsler-Reiss max-stable distribution. Thus, we need first to make sure all the components converges to the multivariate Hüsler-Reiss distribution based on the minimum risk functional. Then, the sample covariance matrix of the threshold exceedances will preserves the independence information as block diagonal matrix. With this sample covariance matrix, we then learn the disconnected clusters by minimising the same objective function as in (Kumar *et al.*, 2019) with spectral constraints. Additionally, as one can regard the extremal correlation as a measure of similarities, we also propose a hierarchical clustering method (Murtagh and Contreras, 2017) based on the estimated extremal correlation matrix to identify multiple independent Hüsler-Reiss Pareto components, which provides another tool for clustering independent components.

This paper is organized as follows: In Section 2, we start by establishing results for r-Pareto processes equipped with marginal exponential tails, where we show that with a new family of risk functional will leads to a r-Pareto processes with the same exponent function. We then introduced the definition of the Hüsler-Reiss graphical models and the extremal covariance matrix, as well as two estimators of the extremal covariance matrix in Section 3. In Section 4, we introduced the spectral graphical lasso method based on the extremal covariance matrix. When we have multiple disconnected graphical components, we assume they corresponds to independent multivariate Pareto distributions. We also discussed the connection between the independence definitions from Engelke *et al.* (2024) and Strokorb (2020), and showed that their definition is not well defined. We establish two different methods to learn the disconnected components, one is based on the spectral constraints and the other is based on hierarchical clustering method using the extremal correlation matrix. In Section 5 and Section 6, we validate our method through simulation studies and two real data applications. We conclude with a discussion in Section 7.

2 Pareto Processes with Exponential Marginal Tail

Multivariate Pareto distributions are a class of distributions that are used to model peaks over thresholds exceedances defined via a homogeneous risk functional $r(\cdot)$, i.e., $r(cx) = cr(x), c > 0$ (Dombry and Ribatet, 2015), when marginally equipped with a Pareto tail. However, when we simply take a logarithmic transformation of the Pareto distributions and assume the marginal distribution has standard exponential tails, it becomes unclear how to define the associated risk functional $r(\cdot)$. Direct mapping the exponential tail back to the Pareto tail by taking exponential transformation make the interpretation and inference awkward as $r(\exp(\cdot))$ is no more a homogeneous function, whereas homogeneity is often a desirable property. In this section, we extend the concept of Pareto processes to the case where marginal tail distribution is assumed to be standard exponential. We borrow the notion in Dombry and Ribatet (2015) and introduce modified theory of Dombry and Ribatet (2015, Theorem 1) by simply taking logarithmic marginal transformation of the random variables, where equivalency between the convergences of the measure of pointwise maxima and the measure of the threshold exceedances associated with the supremum function for random variables with standard exponential tail is established. Let T be a compact metric space and $\mathcal{C} = \mathcal{C}\{T, [-\infty, \infty)\}$, the Banach space of continuous functions from T to $[-\infty, \infty)$, and $\mathcal{C}_0 = \mathcal{C} \backslash \{-\infty\}$.

Proposition 1 (Theorem 1 in Dombry and Ribatet (2015))**.** *Let $X_1, X_2, \dots$ be independent copies of a random process X with samples path in $\mathcal{C}_0$ and standard exponential tail, i.e., $\exists\ u \in \mathbb{R}, \Pr(X(t) - u > x | X(t) > u) = e^{-x}$ for $x > 0, t \in T$. Let $a(n) = \sup\{x \geq 0 : \Pr(\sup_t X(t) \leq x) \leq 1 - 1/n\}$, then, the following statements are equivalent:*

1. *$M_n(t) = \max\{X_1(t), \dots, X_n(t)\} - a(n), t \in T$ converges in distribution to a max-stable random process with exponent measure $\Lambda(\cdot)$ and Gumbel margins as $n \to \infty$, where $\Lambda(u + A) = \exp(-u)\Lambda(A)$ for all measureable set $A \subset \mathcal{C}_0$ and Λ is a continuous measure on $\mathcal{C}_0$.*

2. *$n\Pr\left(X - a(n) \in \cdot\right)$ converges weakly to $\Lambda(\cdot)$ as $n \to \infty$.*

3. *$\Pr\left(X - n \in \cdot | \sup_t X(t) > n\right) \to \Lambda(\cdot \cap \mathcal{C}_{\text{sup}})/\Lambda(\mathcal{C}_{\text{sup}})$ as $n \to \infty$, where $\mathcal{C}_{\text{sup}} = \{x \in \mathcal{C} : \sup_t x(t) > 0\}$.*

In contrast to the homogeneity assumption typically imposed on $r(\cdot)$ for r-Pareto processes with marginal Pareto tails, we instead require that the risk functional $r(\cdot)$ satisfies the following linearity condition:

$$r(x + a) = r(x) + a, a \in \mathbb{R},\ x \in \mathcal{C}_0. \tag{1}$$

This class of risk functional includes, $r(x) = \sup_t x(t), r(x) = \inf_t x(t), r(x) = x(t), t \in T$, and $r(x) = \int_T w(t)x(t)\mathrm{d}t/|T|$, where $w(t) \geq 0$ and $\int_T w(t)\mathrm{d}t = |T|$. The following two theorems establish the equivalency between the convergence of the measure of pointwise maxima and the measure of the threshold exceedances associated with the supremum function for random variables with standard exponential tail, which is a direct extension of Proposition 1 to the case where $r(\cdot)$ satisfies the linearity condition in (1).

Theorem 1. *Assume $r(\cdot)$ satisfies the linearity condition in (1), and let X be a random* *process such that the following weak convergence holds,*

$$\Pr(X - u \in \cdot | r(X) > u) \to \Pr(Y \in \cdot), u \to \infty.$$

Then, the random process Y is either a Pareto process or $\Pr(r(Y) = 0) = 1$. Moreover, if Y is a Pareto process, then the following statements hold and are equivalent:

1. $\Pr(r(Y) > 0) > 0$ *and* $\Pr(Y - u \in \cdot | r(Y) > u) = \Pr(Y \in \cdot)$ *for* $u \geq 0$.

2. $r(Y)$ *has a exponential distribution, i.e.,* $\Pr(r(Y) > u) = e^{-u}$ *for* $u \geq 0$*, and* $r(Y)$ *and* $Y - r(Y)$ *are independent.*

3. $\Pr(r(Y) > 0) = 1$*, and for* $u \geq 0$ *and measureable set* $A \subset \mathcal{C}_r$*, we have*

$$\Pr(Y \in A + u) = \exp(-u)\Pr(Y \in A).$$

The proof of Theorem 1 can be found in Appendix A.1.

Theorem 2. *Suppose the random process X satisfies any statement in Proposition 1 and $r(\cdot)$ satisfies the linearity condition in (1), then we have the following weak convergence:*

$$\Pr(X - u \in \cdot | r(X) > u) \to \Pr(Y \in \cdot), n \to \infty,$$

where Y is a Pareto process with risk functional $r(\cdot)$ as defined in (2).

The proof of above theorem is similar to that of Dombry and Ribatet (2015, Theorem 3), which use the convergence in the second statement of Proposition 1 as a condition to prove the convergence above, then use the results in Theorem 1 to show the limit distribution is a Pareto process defined in (2).

Theorem 2 and 1 combined suggest that, if X is in the max-domain of attraction of a max-stable process with Gumbel margins, then its threshold exceedances over the exceedance region defined via the risk functional, $r(\cdot)$, which satisfy the linearity condition in (1), will converge weakly to a Pareto process Y defined in (2). Different risk functional $r(\cdot)$ yields different Pareto processes but with the same exponent function $\Lambda(\cdot)$, and with such risk functional satisfying (1), the Pareto process Y has the distribution

$$\Pr(Y \in \cdot) = \Lambda(\cdot \cap \mathcal{C}_r)/\Lambda(\mathcal{C}_r),\ \mathcal{C}_r = \{x \in \mathcal{C}_0 : r(x) > 0\}, \tag{2}$$

where $\Lambda(u + A) = \exp(-u)\Lambda(A), u \in \mathbb{R}$. For extremal graphical learning, this allows us to choose certain risk functional such that the inference can be significantly simplified and can be made efficiently. In the next section, we introduce the Hüsler-Reiss graphical models, which is a Hüsler-Reiss Pareto process restricted onto finite dimensions, and the associated extremal covariance matrix.

3 Hus̈ler-Reiss Graphical Model and the Extremal Covariance Matrix

To introduce the Hüsler-Reiss graphical models, we first introduce its corresponding max-stable distributions, which are the essential building block for multivariate Pareto distributions. Max-stable processes are a class of models used for modeling extremes, particularly for block maxima. They serve as the limiting process for componentwise maxima after suitable affine transformation (Resnick, 2008). It has been established that max-stable processes, denoxted as $Z(s)$, with unit Gumbel margins, i.e., $\Pr(Z < z) = \exp\{-\exp(-z)\}$ for $z \in \mathbb{R}$, can be represented using a spectral representation (de Haan, 1984). The spectral representation is given by,

$$Z(s) = \sup_{i=1}^{\infty} R_i + W_i(s) - a(s),\ s \in \mathcal{S},\ R_i \sim \text{PPP}(\exp(-r)),\ \text{for } r \in \mathbb{R}, \tag{3}$$

where $W(s)$ is a spatial process, R_i are the points of a Poisson point process (PPP) of intensity $\exp(-r)$, and $a(s)$ is a constant such that $\text{E}[\exp(W(s) - a(s))] = 1$. The finite dimensional distribution of $Z(s)$ at locations $s_1, \ldots, s_d$ is expressed as

$$\Pr(\boldsymbol{Z} \leq \boldsymbol{z}) = \exp\{-V(\boldsymbol{z})\},\ V(\boldsymbol{z}) = \int_{\mathbb{R}_+} 1 - \Pr(\boldsymbol{W} - \boldsymbol{a} + r\mathbf{1} \leq \boldsymbol{z}) \mathrm{d} \exp(-r), \tag{4}$$

where $V(\boldsymbol{z}) = \Lambda(\{\boldsymbol{x} \in \mathcal{E} : \max_i x_i - z_i > 0\}),\ \mathcal{E} = [-\infty, \infty)^d \backslash \{-\boldsymbol{\infty}\}$, is called the exponent function and Λ is the exponent measure over the support domain $\mathcal{E}$. Suppose now we take $\boldsymbol{W}$ as a random vector and assume $W_1 = 0$ almost surely, then the exponent function becomes

$$V(\boldsymbol{z}) = \int_{-\infty}^{z_1} \exp(-r)\{1 - \Pr(\boldsymbol{W}_{-1} + r\mathbf{1} - \boldsymbol{a}_{-1} \leq \boldsymbol{z}_{-1})\} \mathrm{d}r + \exp(-z_1). \tag{5}$$

and, the associated intensity function is given by

$$\kappa(\boldsymbol{z}) = -\tfrac{\partial^d V(\boldsymbol{z})}{\partial z_1 \ldots \partial z_d} = \exp(-z_1) f_{\boldsymbol{W}_{-1}}(\boldsymbol{z}_{-1} + \boldsymbol{a}_{-1} - z_1 \mathbf{1}), \tag{6}$$

where f_W denotes the density function of the random vector $\boldsymbol{W}$. The results are summarised in the following proposition.

Proposition 2. *For max-stable processes defined in (3), where the constituent process W is taken as a process satisfying condition that $\mathbb{E}[\exp(W(s) - a(s))] = 1$ and $W(s_0) = 0$ almost surely, the exponent function in at locations $s_1, \ldots, s_d$ is given in (5) and corresponding intensity function is given in (6).*

Suppose we have a random vector, X, with unit exponential margins, which belongs to the max-domain attraction of the aforementioned max-stable distribution. In this case, we have $\max_{i=1}^n \boldsymbol{X}_i - \log(n) \xrightarrow{d} \boldsymbol{Z}$, which can be also expressed as the following convergence (Resnick, 2008, See also Theorem 1),

$$n\Pr[\boldsymbol{X} - \log(n) \in \cdot] \to \Lambda(\cdot),\ n \to \infty, \tag{7}$$

where Λ is the exponent measure. Then, we have the weakly convergences, $(X_2 - X_1, X_3 - X_1, \ldots, X_d - X_1) | X_1 > \log(n) \to \boldsymbol{W}_{-1} - \boldsymbol{a}_{-1}$ as $n \to \infty$ (Engelke *et al.*, 2015, for the

Hüsler-Reiss case). This result enables us to translate the inference for extremes to the inference for $\boldsymbol{W}$, and for the Hüsler-Reiss max-stable process, the spatial process $W(s)$ in (3) is taken to be a Gaussian process. Kabluchko *et al.* (2009) suggested that the Hüsler-Reiss max-stable process only depends on the variogram of the Gaussian process, meaning that the constituent Gaussian processes $W(s)$ in the spectral representation are only defined up to Gaussian increments, i.e., $W(s)$ has zero mean and $\mathbb{E}[(W(s_i)-W(s_j))^2] = \gamma_{ij}$, and the normalizing constant $a(s)$ is chosen accordingly. The exponent function of the Hüsler-Reiss distribution is determined by the variogram matrix $\Gamma = (\gamma_{ij})_{i,j=1}^{d}$. However, the variogram matrix alone cannot fully determine the distribution of the constituent Gaussian vector $\boldsymbol{W}$. To make the Gaussian vector $\boldsymbol{W}$ well-defined, one option is to fix a single component of the Gaussian vector to be zero almost surely, as we did in Proposition 2. This allows us to translate the inference for the extremes of $\boldsymbol{X}$ that are in the domain of attraction of $\boldsymbol{Z}$ to the inference for $\boldsymbol{W}$, a Gaussian random vector. It might be reasonable to assume that the Hüsler-Reiss max-stable random vector $\boldsymbol{Z}$ has a conditional independence structure similar to that of the Gaussian random vector $\boldsymbol{W}$. However, this is not true. In fact, conditional independence for $\boldsymbol{Z}$ with a continuous density implies full independence for $\boldsymbol{Z}$ (Papastathopoulos and Strokorb, 2016). For extremes, Engelke and Hitz (2020) introduced the concept of extremal conditional independence and Markov property for the multivariate Pareto random vector, Y, with exceedances region defined as $\mathcal{L}_{\max}^{u}$, where the risk functional $r(\cdot)$ is the maximum, i.e., $\mathcal{L}_{\max}^{u} = \{\boldsymbol{Y} \in \mathcal{E} : r(\boldsymbol{Y}) = \max_i Y_i > u\}$.

The distribution function for multivariate Pareto distributions can be found as

$$\Pr(\boldsymbol{Y} \leq \boldsymbol{y}) = \lim_{n\to\infty} \Pr(\boldsymbol{X}_i - \log(n) \leq \boldsymbol{y} | r(X) > \log(n)) = \frac{\Lambda(\{\boldsymbol{x}\in\mathcal{E}:r(\boldsymbol{x})>0,\ \boldsymbol{x}\leq\boldsymbol{y}\})}{\Lambda(\{\boldsymbol{x}\in\mathcal{E}:r(\boldsymbol{x})>0\})}. \tag{8}$$

The multivariate density function will be given by

$$\frac{\kappa(y)}{\Lambda(\{\boldsymbol{x}\in\mathcal{E}:r(\boldsymbol{x})>0\})}, \tag{9}$$

and we denote the multivariate Pareto distribution as $\boldsymbol{Y} \sim \mathbb{P}_{\mathcal{L}_r^0}$, where $\Lambda(\mathcal{L}_r^0)$ is positive and finite. For Hüsler-Reiss model, the expression of $\kappa(\cdot)$ in (6) is not unique as mentioned earlier, which depends on the definition of the constituent Gaussian random vector $\boldsymbol{W}$ providing the same variogram matrix, Γ. Hentschel *et al.* (2024) used the expression in (6) and derived the intensity function as

Proposition 3 (Proposition 3.4 in Hentschel *et al.* (2024))**.** *The intensity function $\kappa(\cdot)$ for the Hüsler-Reiss process can be expressed in terms of the variogram matrix Γ as*

$$\kappa(\boldsymbol{y}) = (2\pi)^{-(d-1)/2}(d^{-1}|\Theta|_+)^{1/2}\exp\left(-\tfrac{1}{2}\boldsymbol{y}^\top\Theta\boldsymbol{y} - (\tfrac{1}{2d}\Theta\Gamma\mathbf{1} - \tfrac{1}{d}\mathbf{1})^\top\boldsymbol{y} - \tfrac{1}{8d^2}\mathbf{1}^\top(\Gamma\Theta\Gamma + 2\Gamma)\mathbf{1}\right) \tag{10}$$

where Θ is the Moore-Penrose pseudo-inverse of the matrix $\Sigma_e = -1/2(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top)\Gamma(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top)$ and $|\cdot|_+$ denotes the generalized determinant, which is the product of the non-zero eigenvalues of the matrix.

The matrix Σ_e satisfying $\Sigma_e\mathbf{1} = \mathbf{0}$ is a unique positive semi-definite matrix that defines the intensity function of Hüsler-Reiss process since the extremal precision matrix uniquely defines the intensity function. Therefore, we have the following definition for the extremal covariance matrix.

Definition 1 (Extremal covariance matrix)**.** *The matrix Σ_e is called the extremal covariance matrix of the Hüsler-Reiss process if Σ_e is the Moore-Penrose pesudo inverse of the extremal precision matrix. Both of the extremal covariance matrix and the extremal precision matrix are positive semi-definite such that $\Sigma_e\mathbf{1} = \mathbf{0}, \Theta\mathbf{1} = \mathbf{0}$, and define the same intensity function of Hüsler-Reiss process.*

Let $\mathcal{G} = (V, E)$ be a graph, where the set $V = \{1, \dots, d\}$ denotes the nodes and $E \subset V \times V$ denotes the undirected edges between pairs of nodes. For the multivariate Pareto random vector $\boldsymbol{Y} \sim \mathbb{P}_{\mathcal{L}_{\max}^u}, u \in \mathbb{R}$ with standard exponential tail, $\Pr[Y - u > y | Y > u] = e^{-y}, y \geq 0$, extremal conditional independence between ith and jth components is defined (Engelke and Hitz, 2020) as

$$\boldsymbol{Y}^{(k)} := \boldsymbol{Y}|Y_k > u \text{ and } Y_i^{(k)} \perp\!\!\!\perp Y_j^{(k)}|\boldsymbol{Y}_{\setminus\{i,j\}}^{(k)},\ \forall\, k \in V, \tag{11}$$

and we denote the extremal conditional independence as $Y_i \perp\!\!\!\perp_e Y_j|Y_{\setminus\{i,j\}}$. Notice that the above definition is for standard exponential tailed variables with support $\mathcal{L}^0_{\max}$, the original extremal conditional independence is defined when Y has a unit Pareto tailed distribution with support $\mathcal{L}^1_{\max}$ (Engelke and Hitz, 2020) and corresponding multivariate density function is homogeneous of order $-(d+1)$. However, since the risk functional $r(\boldsymbol{x}) = \max_i x_i$ preserves monotonic marginal transformation, such as the logarithmic transformation, the definitions defined over the exponential tail and the Pareto tail are equivalent. The extremal precision matrix Θ in (10) encodes the information of the graphical structure similar as the precision matrix of Gaussian graphical models, where off-diagonal zero entries imply extremal conditional independence. However, Θ is not of full rank, and it has a rank of $d-1$ with null space $\mathbf{1}$, i.e., $\Theta\mathbf{1} = \mathbf{0}$. Wan and Zhou (2025) proposed a complicated estimator for the extremal covariance matrix Σ_e and then used graphical lasso method, termed as extremal graphical lasso, to learn the graphical structure based on the estimated Σ_e. The estimator for Σ_e is formulated as

$$\widehat{\Sigma}^{(1)} := \frac{1}{d}\sum_{k=1}^{d} S^{(k)} - \left(\frac{1}{d^3}\sum_{k=1}^{d} \mathbf{1}^\top S^{(k)}\mathbf{1}\right)\mathbf{1}\mathbf{1}^\top, \tag{12}$$

where $S^{(k)}$ is the empirical sample covariance of $\boldsymbol{Y}^{(k)} - Y_k^{(k)}\mathbf{1}$. It is worth noting that $\widehat{\Sigma}^{(1)}\mathbf{1} = \mathbf{0}$ by design, indicating that the estimator preserves the null space of Σ_e. Another graphical inference method, called EGlearn, proposed by Engelke *et al.* (2025), learns the graphical structure of each component using the Gaussian graphical lasso method based on the sample covariances $S^{(k)}$. The learned sub-graphs are then aggregated into a single graph using majority voting. Additionally, Engelke *et al.* (2025) proposed an estimator for

Γ, termed as empirical variogram matrix and showed that the estimator $\widehat{\Gamma}$ is a consistent estimator for Γ under certain assumptions. Therefore, a second estimator, denoted as $\widehat{\Sigma}^{(2)}$ for Σ_e will be simply replacing Γ with its estimator $\widehat{\Gamma}$, given by

$$\widehat{\Gamma} = \tfrac{1}{d}\sum_{k=1}^{d}\widehat{\Gamma}^{(k)},\ \widehat{\Sigma}^{(2)} = -1/2(I - \tfrac{1}{d}\mathbf{1}\mathbf{1}^\top)\widehat{\Gamma}(I - \tfrac{1}{d}\mathbf{1}\mathbf{1}^\top), \tag{13}$$

where $\widehat{\Gamma}^{(k)}$ is the empirical variogram matrix of $\boldsymbol{Y}^{(k)} - Y_k^{(k)}\mathbf{1}$, and the estimator $\widehat{\Sigma}^{(2)}$ is also used in Röttger *et al.* (2023). Wan and Zhou (2025) conducted a simulation study showing the performance of their extremal graphical lasso method is comparable with the EGlearn method only in certain cases, but worse in high dimensions ($d = 100$) especially when the truth graph is sparse. However, both of the inference methods do not guarantee connectivity.

Wadsworth and Tawn (2014) provided another expression of the intensity function as

$$\begin{aligned}\kappa(\boldsymbol{y}) =&(2\pi)^{-(d-1)/2}|\Sigma|^{-1/2}|\mathbf{1}^\top\Sigma^{-1}\mathbf{1}|^{-1/2}\times \\ &\exp\left(-\tfrac{1}{2}(\boldsymbol{y}+\boldsymbol{a})^\top\left(\Sigma^{-1} - \tfrac{\Sigma^{-1}\mathbf{1}\mathbf{1}^\top\Sigma^{-1}}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}\right)(\boldsymbol{y}+\boldsymbol{a}) - \tfrac{\mathbf{1}^\top\Sigma^{-1}(\boldsymbol{y}+\boldsymbol{a})}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}} + \tfrac{1}{2}(\mathbf{1}^\top\Sigma^{-1}\mathbf{1})^{-1}\right),\end{aligned} \tag{14}$$

where Σ is a well-defined covariance matrix, i.e., Σ is positive definite, of the Gaussian vector $\boldsymbol{W}$ such that $\Gamma = \operatorname{diag}(\Sigma)\mathbf{1}^\top + \mathbf{1}\operatorname{diag}(\Sigma)^\top - 2\Sigma$. As (10) and (14) are equivalent (Kabluchko *et al.*, 2009), we have $\Theta = \Sigma^{-1} - \Sigma^{-1}\mathbf{1}\mathbf{1}^\top\Sigma^{-1}/(\mathbf{1}^\top\Sigma^{-1}\mathbf{1})$, and they encode the same graphical structure. In the next section of the paper, we will explore the relationship between the matrix Θ and Σ, and establish a new expression of the intensity function directly based on the matrix Θ.

Hentschel *et al.* (2024) mentioned that when restricting the support of $\boldsymbol{Y}$ to the space $\{\boldsymbol{y} \in \mathcal{E} : \mathbf{1}^\top\boldsymbol{y} \geq 0\}$, the intensity function $\kappa(\boldsymbol{y})$ is proportional to the density function of a random variable $\boldsymbol{W}' + R'\mathbf{1}$, where W' is a Gaussian random vector with precision matrix Θ such that $\sum_{k=1}^{d} W'_k = 0$ and $R' \sim \operatorname{Exp}(1)$ with $\boldsymbol{W}' \perp\!\!\!\perp R$. Similar results can be also found in Wan (2024), where they proposed to conduct principal component analysis for multivariate Pareto random variables restricted on the hyperplane.

Let the extremal precision matrix Θ has eigenvalues $\lambda_1 = 0, \lambda_2 > 0, \ldots, \lambda_d > 0$ and their corresponding eigenvectors $\boldsymbol{e}_1 = \mathbf{1}/\sqrt{d}, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ so that

$$\Theta = \sum_{i=1}^{d} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top = ABA^\top$$

where $A = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d)$, $B = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, and $A^\top A = I$. The IGMRF of order 1 with linear constraint of order 1, i.e., $\sum_{i=1}^{d} W_i = 0$, can be sampled as $\boldsymbol{W} = \tilde{A}\boldsymbol{E}$, where $\tilde{A} = (\boldsymbol{e}_2, \ldots, \boldsymbol{e}_d)$ and $\boldsymbol{E} \sim \mathcal{N}(\mathbf{0}, \tilde{B} = \operatorname{diag}(\lambda_2^{-1}, \ldots, \lambda_d^{-1}))$. Obviously, we have $\boldsymbol{W}^\top \mathbf{1} = 0$. The spectral representation for the new model is given by

$$\boldsymbol{Z} = \max_{i=1}^{\infty} R_i \mathbf{1} + \boldsymbol{W}_i - \widehat{\boldsymbol{a}}, \; R_i \sim \operatorname{PPP}(\exp(-r)), \text{ for } r \in \mathbb{R}, \tag{15}$$

where $\boldsymbol{W}_i$ are independent copies of $\boldsymbol{W}$ and $\widehat{\boldsymbol{a}} = \operatorname{diag}(\tilde{A}\tilde{B}^{-1}\tilde{A})/2$. Here, we use the $\widehat{\boldsymbol{a}}$ to differentiate from the notation used in (3). The new max-stable model in (15) has intensity function as

$$\begin{aligned} \kappa(\boldsymbol{x}) &= \int_{\mathbb{R}} \exp(-r) 1_{\{\mathbf{1}^\top(x - r\mathbf{1} + \widehat{\boldsymbol{a}}) = 0\}} \phi(\boldsymbol{x} - r\mathbf{1} + \widehat{\boldsymbol{a}}) \mathrm{d}r \qquad (16) \\ &= (2\pi)^{-(d-1)/2} \left(\prod_{i=2}^{d} \lambda_i^{1/2} \right) \exp\left\{ -\tfrac{1}{2}(\boldsymbol{x} + \widehat{\boldsymbol{a}})^\top \Theta (\boldsymbol{x} + \widehat{\boldsymbol{a}}) - \tfrac{(\boldsymbol{x} + \widehat{\boldsymbol{a}})^T \mathbf{1}}{d} \right\}, \boldsymbol{x} \in \mathcal{E} \end{aligned}$$

As for the corresponding multivariate Pareto distribution $\boldsymbol{Y}$ with risk functional $r(\boldsymbol{x}) = 1/d \sum_i x_i$, we have the independence between $1/d \sum_i^d Y_i$ and $\boldsymbol{Y} - \mathbf{1}/d \sum_i^d Y_i$, which can be easily seen from Theorem 1. The term $1/d \sum_i^d Y_i$ can be regarded as the extremeness level, and $\boldsymbol{Y} - \mathbf{1}/d \sum_i^d Y_i$ represents the extremal dependence. Indeed, one can find $\boldsymbol{Y} \sim \mathbb{P}_{\mathcal{L}^u_{\text{avg}}}$, $\mathcal{L}^u_{\text{avg}} == \{\boldsymbol{x} \in \mathcal{E} : 1/d \sum_i x_i > u\}$, can be expressed as,

$$\boldsymbol{Y} = R\mathbf{1} + W - \widehat{\boldsymbol{a}}, R - \widehat{\boldsymbol{a}}^\top \mathbf{1}/d - u \sim \operatorname{Exp}(1), \boldsymbol{W} \sim \operatorname{GMRF}(\Theta), \; \boldsymbol{W}^\top \mathbf{1} = 0, \tag{17}$$

From above representation, to simulate from the multivariate Pareto distribution exactly over the region $\mathcal{L}^u_{\text{avg}}$, we can simulate the extremeness level and the extremal dependence independently as follows: First, we simulate a GMRF with precision matrix Q with the linear constraint $\boldsymbol{W}^\top \mathbf{1} = 0$. Second, we simulate a random variable R from the standard exponential distribution. The resulting random vector, $\boldsymbol{Y} = R\mathbf{1} + \boldsymbol{W} - \widehat{\boldsymbol{a}}$, $\boldsymbol{Y} \in \mathcal{L}^u_{\text{avg}}$, will follow the multivariate Pareto distribution with the intensity function as above. Moreover, if we set $\Sigma^{-1} = \Theta + \tau/d\mathbf{1}\mathbf{1}^\top, \forall\ \tau > 0$ in (14), then, we have $|\Sigma^{-1}|^{1/2}(\mathbf{1}^\top\Sigma^{-1}\mathbf{1})^{-1/2}$ $= (\tau\sum_{i=2}^d \lambda_i)^{1/2}(\tau d)^{-1/2} = d^{-1/2}(\prod_{i=2}^d \lambda_i^{1/2})$ and $\widehat{\boldsymbol{a}} = \boldsymbol{a} - \frac{1}{2D\tau}\mathbf{1}$. Then, the equation in (14) is equivalent to the equation in (16). Therefore, the new model parametrized by the improper precision matrix Θ with exceedance region $\mathcal{L}^0_{\text{avg}}$ and model in (14) with $\Sigma^{-1} = \Theta + \tau/d\mathbf{1}\mathbf{1}^\top$ over the same region $\mathcal{L}^0_{\text{avg}}$ are the same model with intensity function

$$\kappa(\boldsymbol{x}) = (2\pi)^{-(d-1)/2}|\Theta|_+^{-1/2}d^{-1/2}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}+\boldsymbol{a})^\top\Theta(\boldsymbol{x}+\boldsymbol{a}) - \tfrac{\mathbf{1}^\top(\boldsymbol{x}+\boldsymbol{a})}{d} + \tfrac{1}{2\tau d}\right\},$$

which implies that our multivariate Pareto distribution $\boldsymbol{Y}$ is equivalent to the multivariate Pareto distribution defined in Proposition 3 over the region $\mathcal{L}^u_{\text{avg}}$. Alternatively, as $\mathcal{L}^u_{\text{avg}} \subset \mathcal{L}^u_{\max}$, one can simulate samples from the multivariate Pareto distribution over $\mathcal{L}^u_{\text{avg}}$ using rejection sampling by firstly simulating from the multivariate Pareto distribution over $\mathcal{L}^u_{\max}$ and then take the samples that are in $\mathcal{L}^u_{\text{avg}}$.

Theorem 2 suggests that if the random vector $\boldsymbol{X}$ is in the max-domain of the Hüsler-Reiss max-stable distribution, then threshold exceedances, $\boldsymbol{X} - u|r(X) > u$, converges to $\boldsymbol{Y}$ in distribution as $u \to \infty$ with the same intensity function $\kappa(\cdot)$ but different support domain, $\mathcal{L}^0_r$, where $r(\cdot)$ satisfies the linearity condition in (1). Therefore, based on the representation in (17), the graphical structure of the multivariate Pareto distribution, $\boldsymbol{Y} \sim \mathbb{P}_{\mathcal{L}^0_{\max}}$, can be learnt by using the graphical lasso method with the sample covariance matrix for $\boldsymbol{Y} - \mathbf{1}/d\sum_i^d Y_i, \boldsymbol{Y} \in \mathcal{L}^0_{\text{avg}}$, which is summarised in the following theorem.

Theorem 3. *Let $\boldsymbol{Y}$ follows the multivariate Hüsler-Reiss distribution over the region $\mathcal{L}^0_{\max}$, then the sample covariance matrix, $\widehat{\Sigma}^{(3)}$, of $\boldsymbol{Y} - \mathbf{1}/d\sum_i^d Y_i, \boldsymbol{Y} \in \mathcal{L}^0_{avg}$ is a consistent estimator for the covariance matrix $\Sigma^{(3)} = \tilde{A}\tilde{B}\tilde{A}^\top$ such that $\Sigma^{(3)}\mathbf{1} = \mathbf{0}$ and $\Sigma^{(3)}$ is the Moore-Penrose pseudo-inverse of the matrix Θ in (10).*

The proof is simple: $\boldsymbol{Y} - \mathbf{1}/d\sum_i^d Y_i, \boldsymbol{Y} \in \mathcal{L}_{\text{avg}}^u$ follows the Gaussian distribution with covariance $\tilde{A}\tilde{B}\tilde{A}^\top$, which is a direct result of the representation in (17). $\tilde{A}\tilde{B}\tilde{A}^\top$ is the Moore-Penrose pseudo-inverse of Θ. Hence, the sample covariance $\widehat{\Sigma}^{(3)}$ is a consistant estimator of the Moore-Penrose pseudo-inverse of Θ and we have $\widehat{\Sigma}^{(3)}\mathbf{1} = \mathbf{0}$ by design.

Exploring the expression in (14) further, When $\Sigma^{-1} = \Theta + \tau\text{diag}(1, 0, \ldots, 0)$, the intensity function will be given by

$$\kappa(\boldsymbol{x}) = (2\pi)^{-(d-1)/2}|\Sigma|^{-1/2}\tau^{-1/2}\exp\left\{-\tfrac{1}{2}(\boldsymbol{x}+\boldsymbol{a})^\top\Theta(\boldsymbol{x}+\boldsymbol{a}) - (x_1 + a_1) + \tfrac{1}{2\tau}\right\}.$$

Although above expression are not directly useful for learning the extremal graphical model, they provide insights into the relationship between the extremal precision matrix Θ and the proper covariance matrix Σ in (14). Moreover, it is important to note that the covariance matrix Σ is not unique for a given Θ.

4 Structured Graphical Learning and Extremal Independence

The main goal of this paper is using the estimated extremal covariance matrix $\widehat{\Sigma}^{(1)}$, $\widehat{\Sigma}^{(2)}$ and $\widehat{\Sigma}^{(3)}$ to learn graphical structures for multivariate Pareto distribution, which can also guarantee of connectivity. In the case of disconnected components, we should able to interpreted it as fully independence between the disconnected components. For multivariate Gaussian distribution, when assuming the Gaussian density function $f(\cdot)$ is multivariate total positive of order 2 (MTP2), that is

$$f(\boldsymbol{x} \vee \boldsymbol{y})f(\boldsymbol{x} \wedge \boldsymbol{y}) \geq f(\boldsymbol{x})f(\boldsymbol{y}).$$

Then, the Gaussian precision matrix Q has all non-positive off diagonal elements $Q_{ij} \leq 0, i \neq j$. The MTP2 condition is a way imposing positive associations among variables. For Gaussian models, MTP2 condition means all correlations and partial correlations are nonnegative. Röttger *et al.* (2023) proposed a similar concept for multivariate Pareto random vector $\boldsymbol{Y}$ as well, which is called EMTP2, and they call $\boldsymbol{Y}$ is EMTP2 if and only if $\boldsymbol{Y}^{(k)}$ is MTP2 for all $k = 1, \ldots, d$. If $\boldsymbol{Y}$ is EMTP2, then, the precision matrix Θ is a laplacian matrix, i.e., $\Theta \in \mathcal{M} = \{M \in \mathbb{R}^{d\times d} : M\mathbf{1} = \mathbf{0}; M_{ij} < 0, i \neq j\}$. The EMTP2 condition is not restrictive at all as one would expect positive associations for extreme random vectors and there indeed many classic max-stable models that satisfy the EMTP2 condition, including extremal logistic (Tawn, 1990) and extremal Dirichlet distributions (Coles and Tawn, 1991). Indeed, if Θ is a laplacian matrix, the conditional mean of each component in $\boldsymbol{W}$ from (15) defined over the hyperplane $\{\boldsymbol{w} \in \mathcal{E} : \mathbf{1}^\top \boldsymbol{w} = 0\}$ (Rue and Held, 2005) is

$$\mathbb{E}[W_i|\boldsymbol{W}_{-1}] = -\tfrac{1}{\Theta_{ii}} \sum_{j:(i,j)\in E} \Theta_{ij} W_j, \tag{18}$$

which is a weighted average among the neighbours of W_i with positive weights $\Theta_{ij}, (i,j) \in E$ as $-\sum_{j:(i,j)\in E} \Theta_{ij} = \Theta_{ii} > 0$. This allows us to learn the graphical structure of the multivariate Pareto distribution $\boldsymbol{Y}$ from the derivations of $\boldsymbol{W}$ from its overall mean, where the overall mean is specified by the random component R in (15). Intuitively, if a rain storm happens in a region, the graphical dependence structure measures the associations among the derivations of the rain storm at each location from their average over the whole region.

Therefore, from now on, we can reasonably assume Θ is a laplacian matrix of the Hüsler-Reiss graphical model of dimension d, and let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ be its d eigenvalues. It is well studied that the multiplicity of the zero eigenvalue of a laplacian matrix equals to the number of disconnected components in the graph (Chung, 1997). For the extremal precision matrix Θ, we have $\text{rank}(\Theta) = d - 1$ and the multiplicity of the zero eigenvalue is 1, which means the graph is connected (Hentschel *et al.*, 2024; Engelke and Hitz, 2020). Kumar *et al.* (2019) proposed a method to learn the Gaussian graphical model with spectral constraints,

$\mathcal{S}_\lambda$, over the eigenvalues $\lambda_i, 1 = 1, \ldots, d$, where $\mathcal{S}_\lambda$ is specified as

$$\mathcal{S}_\lambda = \{\{\lambda_i = 0\}_{i=1}^m, c_1 \leq \lambda_{m+1} \leq \cdots \leq \lambda_d < c_2\}. \tag{19}$$

The constants c_1 and c_2 are the chosen hyperparameters that represent the minimum and maximum for the eigenvalues, and m is the number of disconnected components in the graph $\mathcal{G}$, which should be 1 for the Hüsler-Reiss graphical model.

When $m > 1$, laplacian matrices with eigenvalues in $\mathcal{S}_\lambda$ are block diagonal matrices with m blocks, and the pesudo-inverse of the laplacian matrix is also a block diagonal matrix with m blocks. For a Gaussian graphical model with such laplacian matrices, the graph will be disconnected with m disconnected components, and disconnected components are independent with each other. The problem of learning the precision matrix Θ giving matrix S can be formulated as maximizing the following objective function (Kumar *et al.*, 2019)

$$\text{maximize}_\Theta \log |\Theta|_\dagger - \text{tr}\,(\Theta S) - \alpha h(\Theta), \tag{20}$$

where $|\cdot|_\dagger$ denotes the generalized determinant, $\text{tr}(\cdot)$ is the trace operator, $h(\Theta)$ is a L_1 norm penalty function on all off-diagonal elements of Θ, and $\alpha > 0$ is a regularization parameter. If we restrict the precision matrix Θ to be a block diagonal laplacian matrix with m blocks, giving the matrix S is also a diagonal block matrix with m blocks. The precision matrix $\widehat{\Theta}$ that maximizes the objective function in (20) will have the same block structure as S. This can be shown by factorizing the term $\text{tr}(\Theta S)$ according to the block structure of S and take $\widehat{\Theta} = \sum_{i=m+1}^{d} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top$. Then, $\boldsymbol{e}_i \boldsymbol{e}_i^\top$ will have the same block structure as S.

Suppose now we have random vector $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)$ with exponential marginal tail, where sets A and B form a partition, and $\boldsymbol{X}$ is in the max-domain attraction of the a max-stable distribution with exponent measure Λ. Then, we call $\boldsymbol{X}_{A_i}$, $i = 1, \ldots, m$, are extremal independent if

$$\lim_{n\to\infty} n^2 \text{Pr}(\boldsymbol{X}_A - \log(n) \in C_A, \boldsymbol{X}_B - \log(n) \in C_B) = \Lambda_A(C_A) \times \Lambda_B(C_B) \tag{21}$$

where $\Lambda_I(\cdot) = \Lambda(\{\boldsymbol{x} \in \mathcal{E} : \boldsymbol{x}_I \in \cdot, \ \cdot \subset \mathcal{E}^I\})$ and $\mathcal{E}^I = [-\infty, \infty]^{|I|} \backslash \{-\boldsymbol{\infty}\}$. With the minimum risk functional, we can have

$$\lim_{u \to \infty} \Pr(\boldsymbol{X}_A - u \in C_A, \boldsymbol{X}_B - u \in C_B | \min_i X_i > u) = \frac{\Lambda_A(C_A \cap \mathcal{L}^0_{\min,A}) \times \Lambda_B(C_B \cap \mathcal{L}^0_{\min,B})}{\Lambda_A(\mathcal{L}^0_{\min,A})\Lambda_B(\mathcal{L}^0_{\min,B})}, \tag{22}$$

where $\mathcal{L}^0_{\min,I} = \{\boldsymbol{x} \in \mathcal{E}^I : \min_i x_i > 0\}$. Thus, we can bring all the components $\boldsymbol{X}_i$ into their Hüsler-Reiss multivariate Pareto distribution with the minimum risk functional. Then, the matrix S that can preserving the extreme independence information of the corresponding extreme random vector $\boldsymbol{Y}_A$ and $\boldsymbol{Y}_B$ as diagonal block structure can be constructed as the sample covariance matrix of $\boldsymbol{X} | \min_i \boldsymbol{X}_i > u$, where u is a large threshold.

This definition of extremal independence is different from the one defined in Engelke *et al.* (2024, Proposition 5.1), where they defined the extremal independence through the infinite exponent measure Λ. They call it extremal independence between components in A and B if for any $\boldsymbol{Y} \sim \mathbb{P}_R$ such that $\Lambda(R) \in (0, \infty)$, we have $\boldsymbol{Y}_A$ and $\boldsymbol{Y}_B$ are independent. They found the extremal independence between $\boldsymbol{Y}_A$ and $\boldsymbol{Y}_B$ only exists when the joint exponent measure put mass on the subspaces only, i.e., $\Lambda(\boldsymbol{x}_A \neq -\boldsymbol{\infty}_A \ \& \ \boldsymbol{x}_B \neq -\boldsymbol{\infty}_B) = 0$, assuming $\Lambda(C)$ is finite over any Borel set, C, that bounds away from the infinity point $\{-\boldsymbol{\infty}\}$. Engelke *et al.* (2024) and Strokorb (2020) showed its equivalency to the traditional extremal independence that is the corresponding max-stable distributions with exponent measure Λ are independent. However, if we use their definition of extremal independence and let $R_{a,\epsilon} = \{\boldsymbol{x} \in \mathcal{E} : e^{x_a} > \epsilon\}$, $\epsilon > 0, a \in A$, $\Lambda(R_{a,\epsilon}) \in (0, \infty)$. Then, we have $\Lambda(R_{a,\epsilon}) = \Lambda_A(R^A_{a,\epsilon}) + \Lambda_B(\mathcal{E}^B)$, where $R^A_{a,\epsilon} = \{\boldsymbol{x} \in \mathcal{E}^A : e^{x_a} > \epsilon\}$, since $R^A_{a,\epsilon}$ and $\mathcal{E}^B$ are the restricted sets of $R_{a,\epsilon}$ onto the subspaces $\mathcal{E}^A$ and $\mathcal{E}^B$ and Λ only has positive mass on the subspaces, i.e., $\{-\boldsymbol{\infty}_A\} \times \mathcal{E}^B \cup \mathcal{E}^A \times \{-\boldsymbol{\infty}_B\}$. Therefore, we have $\Lambda_B(\mathcal{E}^B) \in (0, \infty)$, which implies that marginally $\boldsymbol{Y}_B$ has positive mass at the lower bound of the support, $-\boldsymbol{\infty}_B$. With the same logic, we have $\Lambda_A(\mathcal{E}^A) \in (0, \infty)$, which further implies the exponent measure Λ is a finite measure over $\mathcal{E}$. This result contradicts with the assumption that Λ is an infinite measure.

Instead, use our definition of extremal independence in (21), we can notice that the definition of the joint exponent measure Λ should be defined as a product measure, i.e., $\Lambda(C_A \times C_B) = \Lambda_A(C_A) \times \Lambda_B(C_B)$, whose support domain should be $\mathcal{E}_A \times \mathcal{E}_B$, instead of $\mathcal{E}$. Given the infinite measure Λ and independence between $\boldsymbol{Y}_A$ and $\boldsymbol{Y}_B$, the marginal exponent measures Λ_A and Λ_B should also be infinite measures over their marginal support, otherwise some joint support restriction should be imposed and the joint support domain is not a product space any more. Assuming Λ is an infinite measure, our definition leads to the removal of the mass on the lower bound of the support $-\boldsymbol{\infty}_A$ and $-\boldsymbol{\infty}_B$, meaning there is no mass on any subspaces $\{-\boldsymbol{\infty}_A\} \times \mathcal{E}^B$ and $\mathcal{E}^A \times \{-\boldsymbol{\infty}_B\}$. This is consistent with the independence definition in classic probability theory, and assuming the independence between $\boldsymbol{Y}_A$ and $\boldsymbol{Y}_B$, we should have the marginal exponent measures Λ_A and Λ_B first to define the joint exponent measure Λ. The definition in (21) also extends natually to the case of multiple components, $m > 2$. Unlike the extremal independence definition in Engelke *et al.* (2024), the joint exponent measure should has the property that $\Lambda((C_A+u_A)\times(C_B+u_B)) = \exp(-u_A - u_B)\Lambda(C_A \times C_B)$, where u_A and u_B are constants. One should notice that our definition is not equivalent to the traditional extremal independence defined using max-stable distributions. The intuition behind this is that the convergence of $\boldsymbol{X}$ by taking threshold exceedances or pointwise maxima are different, just like we can have conditional independence for multivaraite Pareto distributions but not for their corresponding max-stable distributions. Further discussions on extremal independence is out of scope of this paper. Next, we will use the sample covariance matrix of $\boldsymbol{X} \,|\, \min_i X_i > u$, where u is a large threshold, together with the structural graphical lasso method introduced later to infer the independent components in the graph $\mathcal{G}$.

To solve the optimisation problem in (20), we can reformulated it as the following minimization problem as in Kumar *et al.* (2019), termed as spectral graphical lasso,

$$\text{minimize}_{\boldsymbol{w},\boldsymbol{\lambda},U} -\log \left|U\text{diag}(\boldsymbol{\lambda})U^\top\right|_\dagger + \text{tr}\left(\mathcal{F}\boldsymbol{w}(\widehat{\Sigma}^{(i)} + \alpha(I - \mathbf{1}\mathbf{1}^\top))\right) + \tfrac{\beta}{2}\|\mathcal{F}\boldsymbol{w} - U\text{diag}(\boldsymbol{\lambda})U^\top\|_F^2 \tag{23}$$

$$\text{subject to } \boldsymbol{\lambda} \in \mathcal{S}_\lambda, \boldsymbol{w} > \mathbf{0}, U^\top U = I,$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathcal{F}$ is a linear operator that transform $\boldsymbol{w} \in \mathbb{R}_+^{d\times(d-1)/2}$ to a laplacian matrix as

$$(\mathcal{F}\boldsymbol{w})_{ij} = \begin{cases} -w_{i+n_j}, & i > j \\ (\mathcal{F}\boldsymbol{w})_{ji}, & i < j \\ -\sum_{i' \neq j}(\mathcal{F}\boldsymbol{w})_{i'j}, & i = j. \end{cases} \tag{24}$$

The linear operator $\mathcal{F}\boldsymbol{w}$ is designed to ensure that $\mathcal{F}\boldsymbol{w} \in \mathcal{M}$. The last term in (23) ensures that $\mathcal{F}\boldsymbol{w}$ converges to the matrix $U\text{diag}(\lambda)U^\top$, where each column of U represents the orthogonal eigenvectors of $\mathcal{F}\boldsymbol{w}$. This allows us to control the eigenvalues of $\mathcal{F}\boldsymbol{w}$ by penalising it towards the desired group structure presented in $U\text{diag}(\lambda)U^\top$ with the penalising parameter β. The first term in (23) is the log-determinant of the precision matrix Θ, and the second term is the trace of the product of $\mathcal{F}\boldsymbol{w}$ and the estimated covariance matrix $\widehat{\Sigma}^{(i)}$, with an additional regularization term that promotes sparsity in $\boldsymbol{w}$ through the parameter α.

When we do clustering, we adopt an adaptive strategy for hyperparameter β inside the optimizer, that is we increase the value of β to enforce connectivity, when the number of zero eigenvalues in $\boldsymbol{\lambda}$ is larger than m, and decrease the value of β when the number of zero eigenvalues in $\boldsymbol{\lambda}$ is smaller than m. By doing so, we can make sure the number of disconnected components in the learnt graph is exactly m. In practice, we first used the adaptive strategy to learn the graph clusters, and then, for each learnt graph cluster, we learn the graph with a mild β, e.g., $\beta = 1$ to ensure fair estimation within each cluster.

Another way of identifying the number of independent components is to use the hierarchical clustering method, which is a well-know method for clustering (Murtagh and Contreras, 2017) in a data-driven approach. The hierarchical clustering method use a dissimilarity matrix to build a cluster tree. A naive way of constructing the dissimilarity matrix is to use the extremal correlation matrix, χ, where χ_{ij} is defined as follows,

$$\chi_{ij} = \lim_{u \to \infty} \Pr(X_i > u | X_j > u). \tag{25}$$

The extremal correlation χ_{ij} measures the extremal dependence between two components, X_i and X_j, and it is bounded between 0 and 1. To estimate the extremal correlation, we can choose a high threshold u and use the empirical conditional probability to approximate it. If $\chi_{ij} = 0$, then X_i and X_j are asymptotical independent, otherwise they are called asymptotical dependent. If X_i and X_j are extremal independent defined in (21), then $\chi_{ij} = 0$. Then, the dissimilarity matrix can be constructed as $D_{ij} = 1 - \chi_{ij}$. The algorithms assume each point is a cluster at the initial stage. Then, it combines the two closest clusters into a new cluster until all points are in one cluster. The distance between two clusters here is chosen to be the complete linkage function, i.e., the maximum D_{ij} over all pairs of points in the two clusters. In this next section, we start with a simulation study to showcase the performance of the proposed spectral graphical lasso method in (23) and the hierarchical clustering method with various threshold u in estimating the extremal correlation matrix χ in (25).

5 Simulation Study

We first examine the performance of the spectral graphical lasso in (23) together with the hierarchical clustering method through simulation on various extreme graphical structures. For this purpose, we randomly generated precision matrices Θ based on Barabśi–Albert (BA) model (Albert and Barabási, 2002), which is an algorithms for generating random graphs with a power-law degree distribution. The BA model contains various real-world graph structures, including World Wide Web, citation network, and social network. It generates graph in a way that, if a new node is added, it is more likely to connect to existing nodes that already have a high degree (connections). The BA model has two parameters, the number of nodes, d, and a degree parameter q, which is the added number of edges when a new node creates. If $q = 1$, then we will have a tree graph structure. Suppose now we have a graph $\mathcal{G} = (V, E)$, the corresponding precision matrix, Θ, is generated by the following:

$$\Theta_{ij} = \begin{cases} -\text{Unif}(0.1, 5), & (i, j) \in E,\ i > j \\ \quad \Theta_{ji}, & \quad i < j \\ -\sum_{i' \neq j} \Theta_{i'j}, & \quad i = j \end{cases} \tag{26}$$

This is the same approach that generates precision matrix as in Engelke *et al.* (2025) except that we use 0.1 instead of 2 in their paper to allow for more flexibility. In our simulation study, we randomly generate connected graphical structure, represented by the extremal precision matrix Θ, with dimension d ranging from 3 to 10 and a random sampled $q = 1$ or $q = 2$. Then, we randomly sample $m = 3, 6, 9$ disconnected components from the previously generated graphs, and generates 100 datasets of $\boldsymbol{Y} \sim \mathbb{P}_{\mathcal{L}^0_{\min}}$ with various replicates n. The number of replicates, n, in each dataset is chosen such that $n/d = 10, 50, 100$. We use the proposed method in (23) and the hierarchical clustering method based on the extremal correlation matrix χ with threshold being the 20%, 50%, 80%, and 90% empirical quantiles. The performance of the proposed methods is evaluated using the adjusted Rand index (ARI) (Rand, 1971; Hubert and Arabie, 1985). Suppose, we have d variables in total, and a true cluster partitioning of the data into m cluster sets, $C_1, \ldots, C_m$. The learnt cluster partitioning consists of m' cluster sets, $C'_1, \ldots, C'_{m'}$. ARI is defined as follows,

$$\begin{aligned}
&\text{ARI} = \tfrac{\text{RI} - \overline{\text{RI}}}{1 - \overline{\text{RI}}}, \quad \text{RI} = \tfrac{a+b}{\binom{d}{2}} \\
&a = \#\left\{(i, j) : 1 \leq i < j \leq d, \exists k_1,\ \exists k_2,\ \text{s.t.},\ i, j \in C_{k_1} \cap C'_{k_2}\right\}, \\
&b = \#\left\{(i, j) : 1 \leq i < j \leq d, \nexists k_1, \nexists k_2,\ \text{s.t.},\ i, j \in C_{k_1} \cap C'_{k_2}\right\}, \\
&\overline{\text{RI}} = \frac{\sum_{i=1}^{m} \binom{\# C_i}{2} \sum_{j=1}^{m'} \binom{\# C'_i}{2}}{\binom{d}{2}},
\end{aligned}$$

where a is the number of pairs of variables that are in the same cluster in both the true and learnt cluster partitioning, and b is the number of pairs of variables that are in different clusters in both the true and learnt cluster partitioning. $\overline{\text{RI}}$ is the expected Rand index (RI) under random clustering. RI measures cluster agreements based on the number of pairs of variables that are consistently partitioned into the same or different clusters, and it is bounded between 0 and 1, where 1 indicates perfect agreement between the true and learnt cluster partitioning, and 0 indicates no agreement between the two cluster partitioning. ARI is a corrected version of RI, which takes into account the randomness of grouping of elements. ARI is bounded between -1 and 1, where 1 indicates perfect agreement between the true and learnt cluster partitioning, 0 indicates the learnt clustering is only as effective as random clustering, and negative values indicate the learnt clustering is worse than random clustering. Figure 1 shows boxplots of adjusted Rand indexes for each case of the pre-mentioned

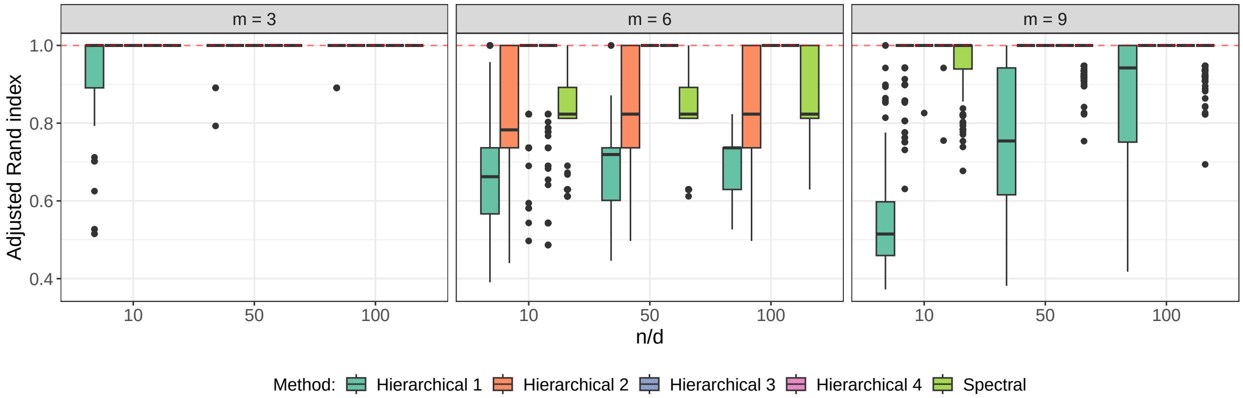

Figure 1: Boxplots of ARI for each case based on 100 simulated datasets using spectral graphical lasso (Spectral) and hierarchical clustering method with threshold u being the 20% (Hierarchical 1), 50% (Hierarchical 2), 80% (Hierarchical 3) and 90% (Hierarchical 4) empirical quantile, where $n/d = 10, 50, 100$ and $m = 3, 6, 9$.

simulation study, indicating the our proposed method based on spectral graphical lasso and the hierarchical clustering method based on the extremal correlation matrix performs well in terms of ARI across all cases. Moreover, the hierarchical clustering method with the 90% empirical quantile consistently yield better results compared with the case with lower threhsold. With reasonable large enough dataset, the hierarchical clustering method with 90% empirical quantile and the spectral graphical lasso provide the best clustering results as

ARIs are close to 1, which means almost perfect pairwise clustering agreements. As the sample size increases, the adjusted Rand index also shows improvements with less variabilities across all cases.

As we have learned the graphical clustering structure, we now proceed to evaluate the performance of the spectral graphical lasso in (23) against the EGlearn method within each cluster. We generate 200 graphs from $\mathrm{BA}(d, q)$ model with a random generated precision matrix Θ, where $d = 5, 10, 20, 50, 100$ and $q = 1, 2$. For each graph and precision matrix, we generated 100 datasets of $\boldsymbol{Y} \sim \mathbb{P}_{\mathcal{L}^0_{\max}}$ with various numbers of replicates, n, which is chosen such that $n/d = 5, 10, 50, 100$. We used the proposed method in (23) with $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$ to learn the extremal graphical structure, and compared it with the EGlearn method. The performance of the proposed method is evaluated using F score, which is defined as

$$\text{F score} = \tfrac{2\#(E \cap E')}{\#E + \#E'}, \tag{27}$$

where E is the true edge set and E' is the learnt edge set. The F score is a measure of prediction accuracy, which considers both the precision and the sensitivity of the prediction to compute the score. The F score is bounded between 0 and 1, where 1 indicates perfect prediction, and the F score is a good measure of performance for the graphical structure learning problem, as it takes both false positives and false negatives cases into account. Figure 2 shows the boxplots of F score for each cases based on different d, q, n and inference method. Results show that the proposed method in (20) with $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$ all yield similar performance as the EGlearn method across all the cases in terms of the F score. Moreover, as n/d increases, the F score also increases and shows less variability.

As the F scores for all the four methods are computed based on each dataset we simulated, the F scores are naturally paired with each other across the four methods we evaluated here. One would expect different precision matrix Θ, which encodes the graph structure, and different sample size will have major impact on the F score. Therefore, we also evaluate the

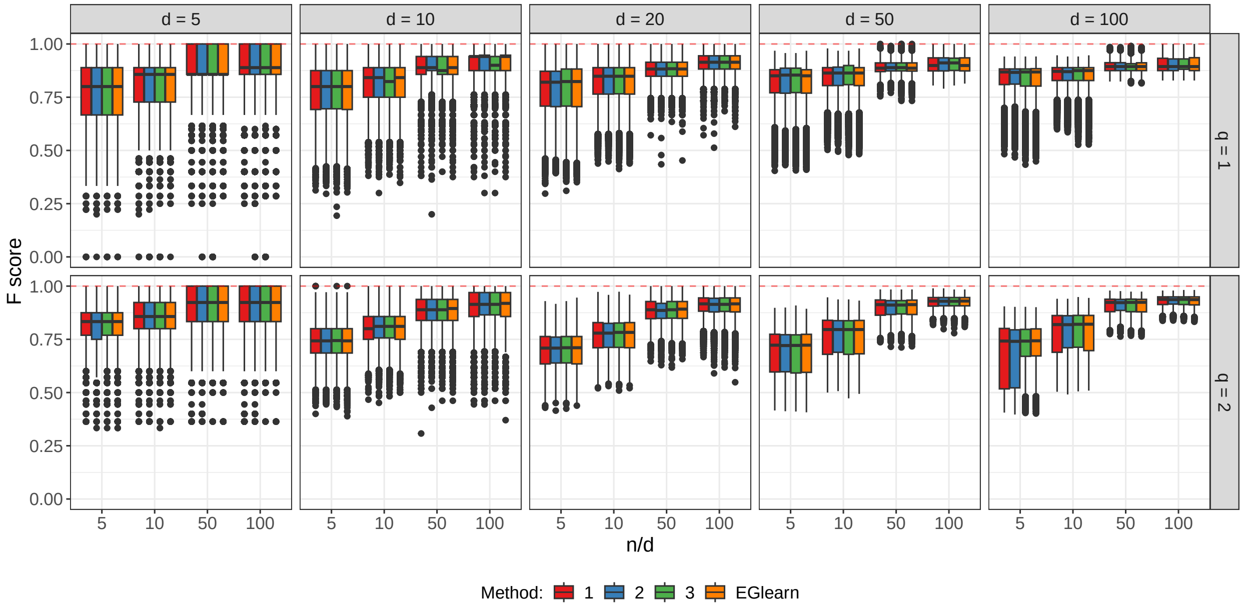

Figure 2: Boxplots of F score for each case based on 200 graphs and 100 simulated datasets for each graph using methods in (20) with $\widehat{\Sigma}^{(1)}$ (red), $\widehat{\Sigma}^{(2)}$ (blue) and $\widehat{\Sigma}^{(3)}$ (green) and the EGlearn method (orange), where $n/d = 5, 10, 20, 50, 100$, $q = 1, 2$, and $d = 5, 10, 20, 50, 100$.

performance differences between our methods and the EGlearn method given a particular precision matrix Θ and sample size. To do this, we conduct a paired t-test on the F scores of the proposed method using $\widehat{\Sigma}^{(i)}$ and the EGlearn method for each $i = 1, 2, 3$. The alternative hypothesis is that the F score of the proposed method using $\widehat{\Sigma}^{(i)}$ is greater than the F score of the EGlearn method. The significance level is set at 0.05. The results are shown in Figure 3. Our proposed method outperforms the EGlearn method in most cases, especially when $q = 2$ with high dimensions, and $q = 1$ with lower dimensions and smaller sample size. With smaller sample size, the EGlearn method performs not as good as our method. In lower dimensional cases, our proposed method using $\widehat{\Sigma}^{(2)}$ and $\widehat{\Sigma}^{(3)}$ performs better than using $\widehat{\Sigma}^{(1)}$. In addition, we also conduct similar paired t-test with alternative hypothesis that the F score of the proposed method using $\widehat{\Sigma}^{(i)}$ is less than the F score of the EGlearn method. The results are shown in Figure 4. The EGlearn method outperforms our methods only when the graph structure is a tree, i.e., $q = 1$, and the sample size is large. As we have

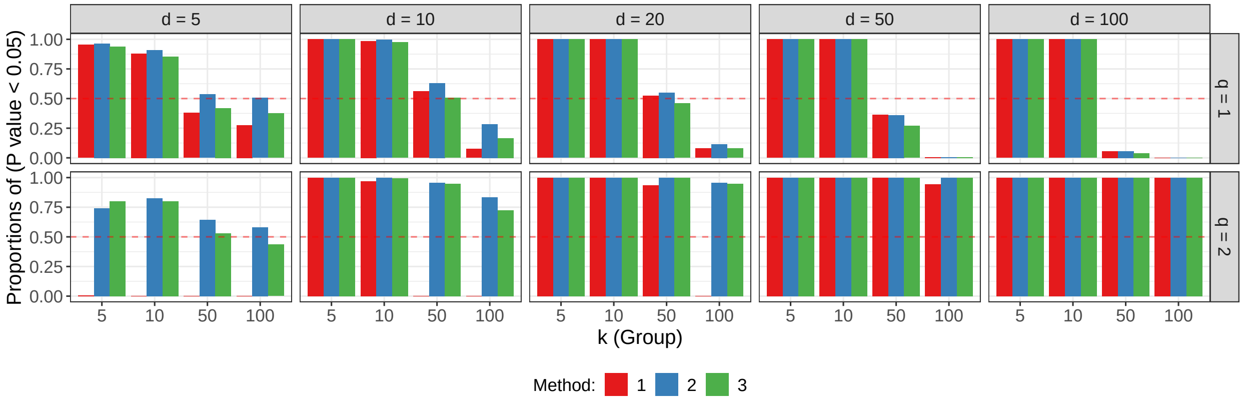

Figure 3: Barplots of the proportion of the F score of the proposed method using $\widehat{\Sigma}^{(i)}$ is greater than the F score of the EGlearn method for each case based on 200 graphs and 100 simulated datasets for each graph using confidence level, 0.05, where $n/d = 5, 10, 20, 50, 100$, $q = 1, 2$, and $d = 5, 10, 20, 50, 100$.

compared the performances of the four methods, one may ask how much the performance differences between our methods and the EGlearn method are. To answer this question, we compute the average difference in 100% between the F score of simulation results using method based on $\widehat{\Sigma}^{(i)}$ and the EGlearn method in each cases, where $i = 1, 2, 3$. The results are shown in Table 1. As expected from the comparison using above boxplots, Table 1 shows

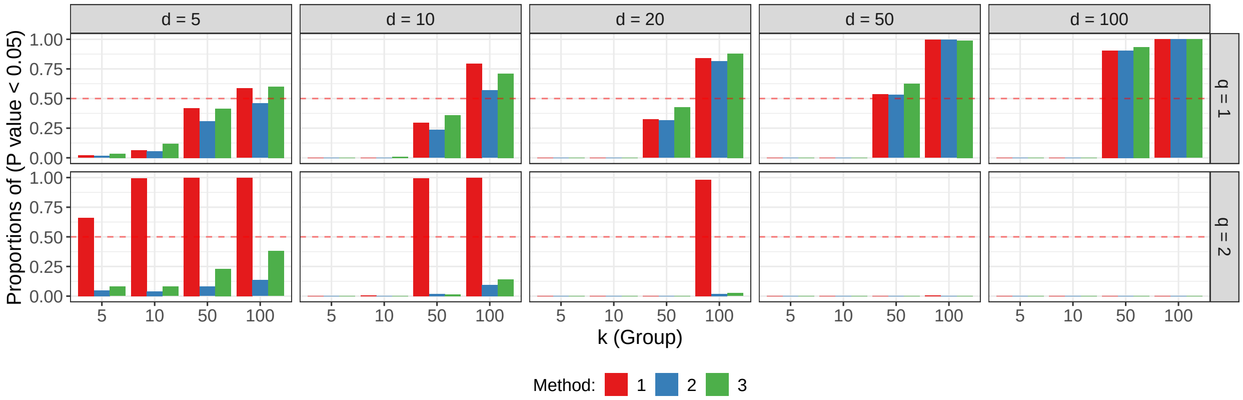

Figure 4: Barplots of the proportion of the F score of the proposed method using $\widehat{\Sigma}^{(i)}$ is less than the F score of the EGlearn method for each case based on 200 graphs and 100 simulated datasets for each graph using confidence level, 0.05, where $n/d = 5, 10, 20, 50, 100$, $q = 1, 2$, $d = 5, 10, 20, 50, 100$, $i = 1$(red),2(blue),3(green). The red dashed line indicates 50% level.

our methods obtain significant performance gains against the EGlearn method when either sample size is limited ($n/d < 10$) or dimension is high ($d > 10$), especially for the case $q = 2$. The average F score differences can be as high as 36.35 out of 100 when $d = 100, n/d = 5$. However, as sample size increases, the advantage of our method over the EGlearn method decreases. In cases when the EGlearn method outperforms our method as indicated by Figure 4, the average F score differences is considered small (less than 10%). Therefore, in summary, our method outperforms the EGlearn method significantly when the sample size is relatively small or dimension is relatively high. The proposed method using $\widehat{\Sigma}^{(3)}$ obtains a larger performance gains than using $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$, which is consistent with the results shown in Figure 3 and Figure 4. A byproduct of the estimators $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$ is that they can be used to estimate the Γ using the transformation $\Gamma = \mathrm{diag}(\Sigma)\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(\Sigma)^\top - 2\Sigma$ by replacing the matrix with corresponding estimates. We can use the mean squared error (MSE) of the estimated Γ to evaluate the performance of the estimators. The MSE is defined as average of the squared differences between every entries of the estimated Γ and the true

Table 1: Average difference in F scores (×100) using method based on $\widehat{\Sigma}^{(i)}$ and the EGlearn method in each cases, where $i = 1, 2, 3$.

		$q = 1$					$q = 2$				
$n/d\backslash d$		5	10	20	50	100	5	10	20	50	100
5	$\widehat{\Sigma}^{(1)}$	14.56	19.29	23.35	23.89	20.13	-3.16	5.53	11.69	21.04	26.58
	$\widehat{\Sigma}^{(2)}$	15.79	20.40	23.88	24.00	20.14	2.39	6.43	12.28	21.34	26.72
	$\widehat{\Sigma}^{(3)}$	16.29	22.40	25.84	25.04	20.72	3.19	10.52	18.44	29.48	36.35
10	$\widehat{\Sigma}^{(1)}$	11.01	14.32	18.00	17.99	14.40	-6.55	4.10	11.28	19.89	24.24
	$\widehat{\Sigma}^{(2)}$	12.41	15.23	18.47	18.14	14.41	2.32	5.31	11.68	20.29	24.41
	$\widehat{\Sigma}^{(3)}$	12.04	15.91	19.16	18.37	14.44	2.83	8.89	17.03	26.93	31.40
50	$\widehat{\Sigma}^{(1)}$	-1.63	0.94	1.37	-0.62	-3.29	-17.00	-6.72	3.42	8.14	8.46
	$\widehat{\Sigma}^{(2)}$	1.96	2.52	1.65	-0.57	-3.28	0.56	3.64	7.20	9.12	8.84
	$\widehat{\Sigma}^{(3)}$	-0.62	0.46	0.37	-1.35	-3.77	0.10	5.13	9.37	11.01	10.47
100	$\widehat{\Sigma}^{(1)}$	-6.42	-4.95	-4.60	-5.91	-7.64	-19.99	-13.98	-4.35	2.51	4.67
	$\widehat{\Sigma}^{(2)}$	-2.48	-2.52	-4.29	-5.89	-7.64	0.11	2.13	3.82	4.54	5.24
	$\widehat{\Sigma}^{(3)}$	-5.74	-5.27	-5.90	-6.84	-8.19	-1.18	2.01	4.03	4.78	5.68

Γ. Table 2 shows the results of the average MSE of the estimated Γ using the estimators $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$ and the estimated laplacian matrix returned by the optimizer in (20) using $\widehat{\Sigma}^{(3)}$, denoted by $\mathcal{F}\boldsymbol{w}_0$. Notice that the estimated variogram matrix using $\widehat{\Sigma}^{(2)}$ will produce the same estimator as the $\widehat{\Gamma}$ in (13). The results show that the proposed method in (20) with $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$ all yield similar performance as the EGlearn method across all the cases in terms of MSE. Surprisingly, when the graph structure is a tree, i.e, (q=1), the MSE is significantly larger than the MSE for the case $q = 2$. When $q = 2$, the mse is much smaller. Our results also show that the proposed method in (20) with $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$ all yield similar performance as the EGlearn method across all the cases in terms of computational time. The computational time increases as d increases. Though EGlearn method needs to learn d sub-graphs, our proposed inference method has much more optimization constraints, $\mathcal{S}_\lambda$,

Table 2: Mean squared error of the estimated Γ averaged across the 200 simulated graphs and 100 datasets for each graph.

		$q = 1$					$q = 2$				
$n/d\backslash d$		5	10	20	50	100	5	10	20	50	100
5	$\widehat{\Sigma}^{(3)}$	2.41	4.15	6.90	9.33	11.14	0.15	0.11	0.07	0.06	0.05
	$\mathcal{F}\boldsymbol{w}_0$	3.06	4.47	7.73	7.71	9.72	0.10	0.07	0.07	0.06	0.05
	$\widehat{\Sigma}^{(2)}$	3.52	4.83	5.70	10.55	10.77	0.13	0.07	0.06	0.06	0.05
	$\widehat{\Sigma}^{(1)}$	2.98	4.76	6.71	10.07	10.19	0.09	0.07	0.05	0.05	0.05
10	$\widehat{\Sigma}^{(3)}$	2.83	3.99	6.70	9.15	11.00	0.19	0.12	0.08	0.08	0.06
	$\mathcal{F}\boldsymbol{w}_0$	2.94	4.36	7.45	7.59	9.54	0.11	0.07	0.08	0.07	0.06
	$\widehat{\Sigma}^{(2)}$	3.36	4.59	5.49	10.30	10.57	0.12	0.06	0.07	0.06	0.06
	$\widehat{\Sigma}^{(1)}$	3.45	4.62	6.66	9.91	9.98	0.07	0.08	0.06	0.06	0.06
50	$\widehat{\Sigma}^{(3)}$	2.70	3.89	6.52	9.01	10.87	0.22	0.13	0.08	0.10	0.07
	$\mathcal{F}\boldsymbol{w}_0$	2.78	4.27	7.28	7.45	9.41	0.10	0.07	0.07	0.09	0.06
	$\widehat{\Sigma}^{(2)}$	3.20	4.42	5.31	10.22	10.43	0.17	0.06	0.07	0.07	0.07
	$\widehat{\Sigma}^{(1)}$	3.24	4.43	6.51	9.81	9.83	0.10	0.08	0.05	0.05	0.06
100	$\widehat{\Sigma}^{(3)}$	2.69	3.86	6.51	9.21	10.98	0.23	0.14	0.07	0.09	0.07
	$\mathcal{F}\boldsymbol{w}_0$	2.74	4.25	7.23	7.45	9.37	0.09	0.07	0.07	0.09	0.06
	$\widehat{\Sigma}^{(2)}$	3.19	4.39	5.27	10.15	10.36	0.17	0.04	0.06	0.08	0.07
	$\widehat{\Sigma}^{(1)}$	3.22	4.41	6.51	9.63	9.77	0.09	0.07	0.05	0.05	0.06

and also need to update the orthogonal matrix U via singular value decomposition of matrix $\mathcal{F}\boldsymbol{w}$ in each iteration in (20). However, our method can still be as computationally efficient, and thus can be used to learn the graphical structure of multivariate Pareto distribution with high dimensions.

6 Applications

In our application, we use two real datasets including one that has been demonstrated in Engelke and Hitz (2020); Hentschel *et al.* (2024), to illustrate our proposed method, which is the Danube river discharge network and is initially studied by Asadi *et al.* (2015). The second dataset is the stock index in major stock markets around the globe. The Danube river network consists of 31 nodes (stations) with daily 428 samples spanning from 1960 to 2010 after being declustered to remove the temporal clusters. The river flow network is shown in Figure 5 at the top left with edge thicknesses indicating the average flow volume between nodes and arrows indicating the flow directions. We first transform the data to the standard exponential margins, and select the threshold, u, to be the 80% quantile in $\mathcal{L}^u_{\min}$ to identify the clusters as in (22) for spectral graphical lasso method and also in (25) for hierarchical clustering method. Once we learnt the clusters, we use 80% empirical quantiles as u within the estimator $\widehat{\Sigma}^{(3)}$ to learn the individual graph within each connected graphical component. We set the number of clusters to be $m = 1, 2, \ldots, 5$, and we plot the heatmap of the estimated χ matrix using 80% empirical quantile as the threshold in (25), together with the learnt clustering structured in Figure 5 shown as colored bars for the rows (using hierarchical clustering method) and columns (using spectral graphical lasso). As $m = 2$, these two methods yield the same clustering structure. For $m = 3, 4, 5$, the RI index between the two estimated clustering structure is maximized at $m = 4$, which is 0.51. The hierarchical clustering method tends to cluster station 1, 13, 14, 15, 16, 17, 18, and 19 together as a single cluster but not with station 2, while the spectral graphical lasso method tends to cluster station 1, 13, 28,29,30,31 together as a single cluster. According to the original river flow network, we found the spectral graphical lasso method yileds more interpretable clustering structuer as stations 1, 13, 28, 29, 30, and 31 are all connected and can be regarded as a single river branch, while stations 1, 13, 14, 15, 16, 17, 18, and 19 are disconnected without station 2. The estimated graph using the spectral graphical lasso method is shown in the second and third column of Figure 5 with colors indicating different clusters for $m = 1, 2, 3, 4$. The learnt graph largly ensemble the original river flow network, and the disconnected components are the points that are far away from the main river network in the upper stream with small flow volume. The major river branches are identified as a single cluster in most cases, and the most right river branch is identified as a cluster by itself in most cases, which is consistent with the fact that its flow does not contribute to the river network on its left side.

The stock markets data contains stock index in 16 major stock exchanges around the world, including Dow Jones Index (DJI) and SP500, which are the two most important stock indexes in the US (denoted by NY: SP500 and NY: DJI), Nikkei225 (N225) in Japan, FTSE100 in the UK, DAX in Germany, CAC40 in France, HSI in Hong Kong, ASX 200 in Australia and so on as shown in Figure 6. The dataset is obtained from Yahoo finance and contains daily logarithmic return from 2000 to March in 2025. For each index, we fit a time series model, ARIMA(2,1,2), to remove the trend and auto-correlation from the data. Then, we obtain the fitted residuals as anomalies. We then transform the anomalies to the standard exponential margins, and select the threshold, u, to be the 90% empirical quantiles when estimate the χ matrix and implement spectral graphical lasso method for clustering as well as doing individual graph learning using spectral graphical lasso method.

The results are shown in Figure 6 as a circle according to their timezones with different colors representing different clusters. Noticed that we moved the trading date of the North

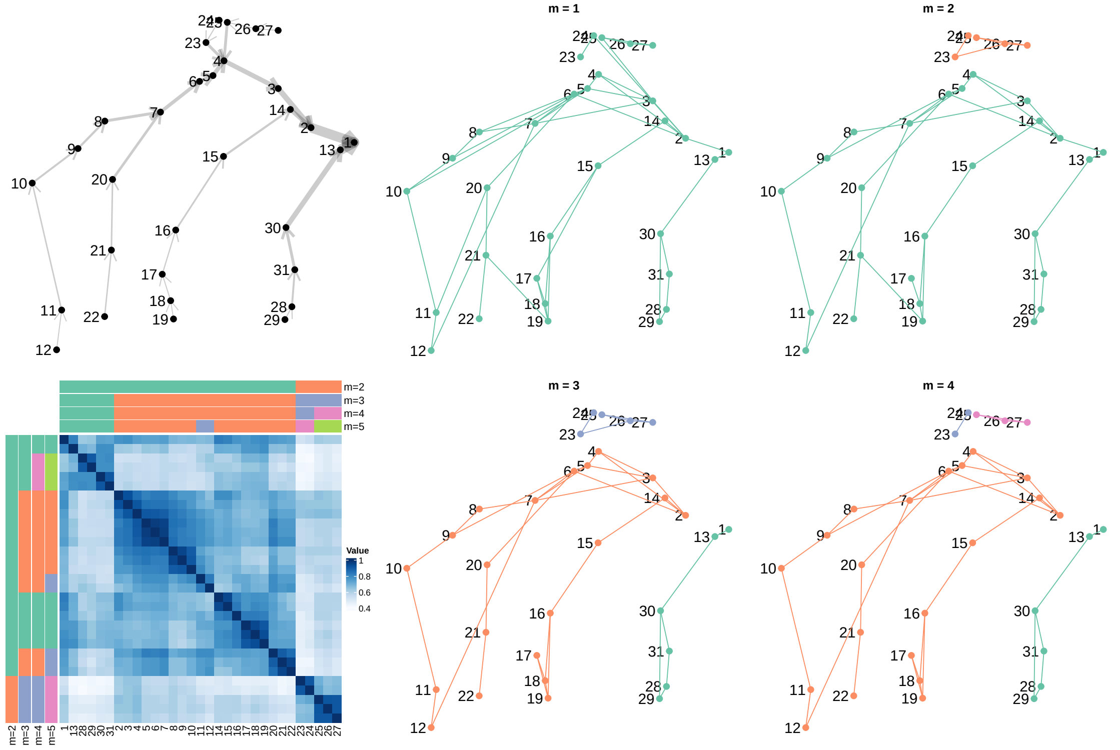

Figure 5: Estimated graph with number of clusters $m = 1, 2, 3, 4$ from the top left panel to the bottom right panel in the second and third columns with colors indicate different clusters within plot. The original river flow network is shown on the top left with line widths indicating the average flow volume between the two nodes and arrows indicating the flow directions.The heatmap at the bottom left shows the estimated χ matrix with the clustering structure shown as colored bars for the rows (using hierarchical clustering method) and columns (using spectral graphical lasso).

America forward by 1 day to match the trading date of the Asia and Europe. We also plot the heatmap of the estimated χ matrix together with the colored bar representing the clusters similar as in Figure 5. The clusters patterns are more presented in the heatmap than in the Danube river data as two methods yield same clustering structures up to $m = 4$. The two Chinese stock indexes (Shenzhen and Shanghai) are always clustered together showing unique market characteristics that are different from the rest developed markets. The (North and South) American stock indexes are clustered together since they are all traded in the same timezone and absorbing similar market shocks at the same time. The developed Asian stock indexes are also clustered together as they are in similar timezones, and have similar market characteristics. The learnt graphical structure seem to suggest the trading timezone can be regarded as the major common factor for the joint extreme market fluctuations, though this has to be further investigated.

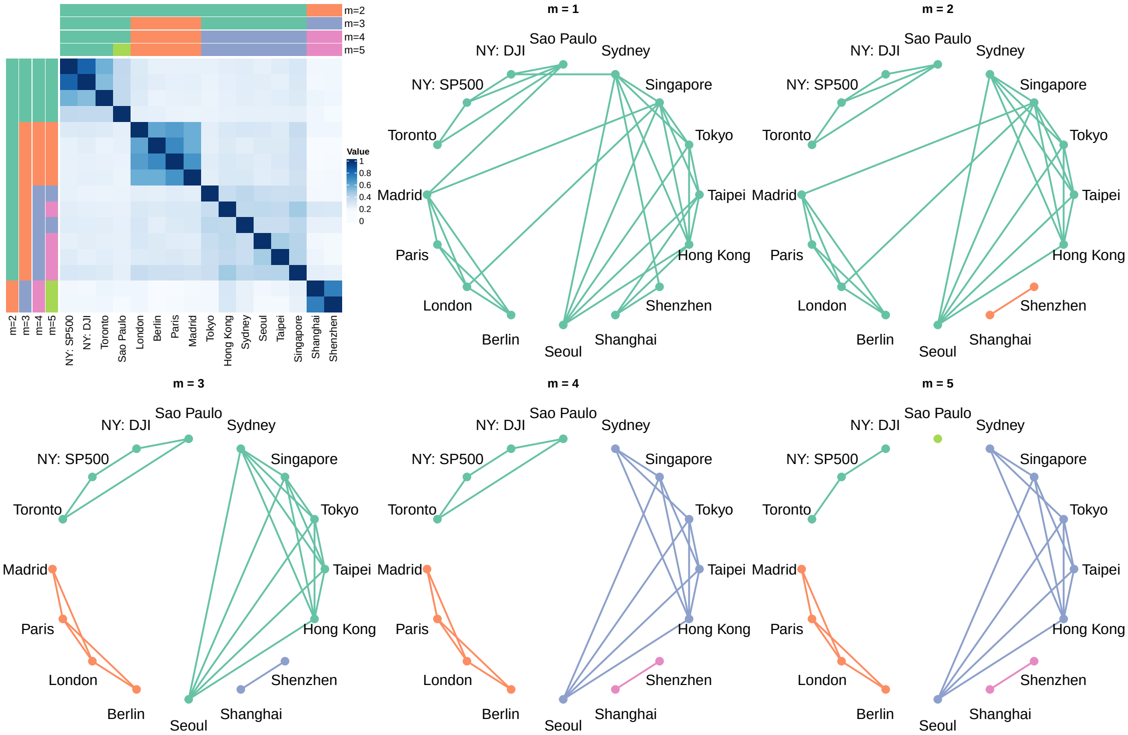

Figure 6: The estimated graphs with number of clusters $m = 1, 2, 3, 4$ from the second panel on the top to the right panel on the bottom. The colors of the nodes and edges indicate different clusters within each graph. The first panel on the top left shows the heatmap of the estimated χ matrix with the clustering structure shown as colored bars for the rows (using hierarchical clustering method) and columns (using spectral graphical lasso).

7 Conclusion

In this paper, we proposed a new method to learn the graphical structure of multivariate Hüsler-Reiss Pareto distribution under the assumption of EMTP2. We also justifies the usage of the proposed method as a clustering method for the Hüsler-Reiss multivariate Pareto distribution. To use the method, we propose three extremal covariance estimator, $\widehat{\Sigma}^{(i)}, i = 1, 2, 3$, where $\widehat{\Sigma}^{(3)}$ is newly proposed, $\widehat{\Sigma}^{(1)}$ is originated from Wan and Zhou (2025), and $\widehat{\Sigma}^{(2)}$ is used in Röttger *et al.* (2023). While those estimators are established based on different support domains, i.e., $\mathcal{L}^0_{\max}$ and $\mathcal{L}_{\mathrm{avg}^0}$, we showed that as long as these restricted set is build based on the risk functional satisfying the linearity condition in (1), then the corresponding multivariate Pareto distribution will have a exponent measure with the same extreme precision matrix. The proposed method is computationally efficient and can be used to learn the graphical structure of multivariate Pareto distribution with high dimensions when comparing with the state of art method EGlearn (Engelke *et al.*, 2025) for learning extreme graphical structure. Our methods outperforms the EGlearn method significantly when the sample size is small or dimension is high.

For clustering, we define the extremal independence based on the classic independence definition, as contract to the extremal independence introduced by (Engelke *et al.*, 2024), where we showed their definition is contradictory and one need to remove the mass of the exponent measure Λ on the subspaces instead of only putting mass on the subspaces. To learn the independent multivariate Pareto components, we proposed two method, one is a hierarchical clustering method using the empirical extremal correlation matrix, and another is the spectral graphical lasso. Our method can consistently recover the true clustering structure as demonstrated by the simulation study. The proposed method is also applied to two real datasets, including the Danube river network and stock markets network, to illustrate its performance. The learnt clustered graphical structure is meaningful and can be sensibly interpreted. In our future work, we could investigate the theoretical asymptotical properties of the proposed method and explore the possibility of extending the method to other multivariate Pareto distributions. Also, as pointed out earlier, the two definition of extremal independence built upon the max-stable distribution (the traditional definition) and the definition based on the infinite exponent measure through the multivariate Pareto distribution, should be also further investigated. Possible way to look at it is through the convergence of the distribution of $\boldsymbol{X}$ towards its extreme limit, whether via pointwise maxima or the threshold exceedance.

References

Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of modern physics* **74**, 47.

Asadi, P., Davison, A. C. and Engelke, S. (2015) Extremes on river networks. *Annals of Applied Statistics* **9**, 2023–2050.

Chung, F. R. (1997) *Spectral graph theory*. Volume 92. American Mathematical Soc.

Coles, S. G. and Tawn, J. A. (1991) Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **53**, 377–392.

Dombry, C. and Ribatet, M. (2015) Functional regular variations, Pareto processes and peaks over threshold. *Statistics and Its Interface* **8**, 9–17.

Engelke, S. and Hitz, A. S. (2020) Graphical models for extremes (with Discussion). *Journal of the Royal Statistical Society: Series B* **82**, 871–932.

Engelke, S., Ivanovs, J. and Strokorb, K. (2024) Graphical models for infinite measures with applications to extremes. *arXiv preprint arXiv:2211.15769* .

Engelke, S., Lalancette, M. and Volgushev, S. (2025) Learning extremal graphical structures in high dimensions. *arXiv preprint arXiv:2111.00840* .

Engelke, S., Malinowski, A., Kabluchko, Z. and Schlather, M. (2015) Estimation of Huesler–Reiss distributions and Brown–Resnick processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**, 239–265.

de Haan, L. (1984) A spectral representation for max-stable processes. *Annals of Probability* **12**, 1194–1204.

Hentschel, M., Engelke, S. and Segers, J. (2024) Statistical inference for hüsler–reiss graphical models through matrix completions. *Journal of the American Statistical Association* pp. 1–25.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification* **2**, 193–218.

Kabluchko, Z., Schlather, M. and de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. *Annals of Probability* **37**, 2042–2065.

Kumar, S., Ying, J., de Miranda Cardoso, J. V. and Palomar, D. (2019) Structured graph learning via laplacian spectral constraints. *Advances in neural information processing systems* **32**.

Lauritzen, S., Uhler, C. and Zwiernik, P. (2019) Maximum likelihood estimation in gaussian models under total positivity. *Annals of Statistics* **47**, 1835–1863.

Murtagh, F. and Contreras, P. (2017) Algorithms for hierarchical clustering: an overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **7**, e1219.

Papastathopoulos, I. and Strokorb, K. (2016) Conditional independence among max-stable laws. *Statistics & Probability Letters* **108**, 9–15.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Resnick, S. I. (2008) *Extreme values, regular variation, and point processes.* Volume 4. Springer Science & Business Media.

Röttger, F., Engelke, S. and Zwiernik, P. (2023) Total positivity in multivariate extremes. *The Annals of Statistics* **51**, 962–1004.

Rue, H. and Held, L. (2005) Gaussian Markov Random Fields: Theory and Applications. In *Monographs on Statistics and Applied Probability*, volume 104. London: Chapman & Hall.

Strokorb, K. (2020) Contribution to the Discussion of the paper 'Graphical models for extremes' by S. Engelke and A. Hitz. *Journal of the Royal Statistical Society: Series B* **82**, 912–913.

Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika* **77**, 245–253.

Wadsworth, J. L. and Tawn, J. A. (2014) Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101**, 1–15.

Wan, P. (2024) Characterizing extremal dependence on a hyperplane. *arXiv preprint arXiv:2411.00573* .

Wan, P. and Zhou, C. (2025) Graphical lasso for extremes. *arXiv preprint arXiv:2307.15004* .

Ying, J., de Miranda Cardoso, J. V. and Palomar, D. (2020) Nonconvex sparse graph learning under laplacian constrained graphical model. *Advances in Neural Information Processing Systems* **33**, 7101–7113.

Ying, J., de Miranda Cardoso, J. V. and Palomar, D. (2021) Minimax estimation of laplacian constrained precision matrices. In *International Conference on Artificial Intelligence and Statistics*, pp. 3736–3744.

A Proofs of Theorems and Lemmas

A.1 Proof of Theorem 1

Proof. We begin the proof by showing the first statement holds given Y is a Pareto process defined in (2), followed by showing all the statements are equivalent. First, we have $\Pr(r(Y) > 0) = \Lambda(\{x \in \mathcal{C}_0 : r(x) > 0\} \cap \mathcal{C}_r)/\Lambda(\mathcal{C}_r) = 1 > 0$. Let A be a measurable set in $\mathcal{C}_0$, then, we have

$$\begin{aligned}\Pr(Y - u \in A | r(Y) > u) &= \Lambda(\{x \in \mathcal{C}_0 : r(x) > u, x \in A + u\})/\Lambda(\{x \in \mathcal{C}_0 : r(x) > u\})\\ &= \Lambda(u + \{x \in \mathcal{C}_0 : r(x) > 0, x \in A\})/\Lambda(u + \{x \in \mathcal{C}_0 : r(x) > 0\})\\ &= \Pr(Y \in A)\end{aligned}$$

. Next, we show the first statement implies the second statement. With $A = \{x \in \mathcal{C}_0 : r(x) > v\}, v \geq 0$, we have $\Pr(Y - u \in A | r(Y) > u) = \Pr(r(Y) > v + u | r(Y) > u) = \Pr(r(Y) > v)$. Thus, $\Pr(r(Y) > v + u) = \Pr(r(Y) > u)\Pr(r(Y) > v)$, which leads to $\Pr(r(Y) > u) = \exp(-u), u \geq 0$. To prove $r(Y)$ and $Y - r(Y)$ are independent, we let $A = \{x \in \mathcal{C}_0 : x - r(x) \in B\}$, and we have

$$\begin{aligned}\Pr(Y - r(Y) \in B, r(Y) > u) &= \Pr(Y \in A, r(Y) > u)\\ &= \Pr(Y - u \in A, r(Y) > u)\\ &= \Pr(Y \in A)\Pr(r(Y) > u)\\ &= \Pr(Y - r(Y) \in B)\Pr(r(Y) > u).\end{aligned}$$

To prove the equivalency between the second and third statement, we define the set $A_{v,B} = \{x \in \mathcal{C}_0 : r(x) \geq v, x - r(x) \in B\}$, where $v \geq 0$ and $B \in \{x \in \mathcal{C}_0 : r(x) = 0\}$ measurable. We have $\Pr(Y \in A_{v,B}) = \Pr(r(Y) \geq v)\Pr(Y - r(Y) \in B) = \exp(-v)\Pr(Y - r(Y) \in B)$. Notice that, $u + A_{v,B} = A_{v+u,B}$, we have, $\Pr(Y \in u + A_{v,B}) = \exp(-u)\Pr(Y \in A_{v,B})$. The sets $A_{v,B}$ forms a $\pi-$system, and hence, the condition $\Pr(Y \in u + A) = \exp(-u)\Pr(Y \in A)$ holds for all measurable set A. Now, it remains to prove the equivalency between the third and the first statement. Let $A \in \mathcal{C}_0$ be a measurable set and $u \geq 0$, then, we have

$$\begin{aligned}\Pr(Y - u \in A | r(Y) > u) &= \Pr(Y - u \in A, r(Y) > u)/\Pr(r(Y) > u)\\ &= \exp(-u)\Pr(Y \in A, r(Y) > 0)/\exp(-u)\Pr(r(Y) > 0) = \Pr(Y \in A).\end{aligned}$$

It remains to show that if

$$\Pr(X - u \in \cdot | r(X) > u) \to \Pr(Y \in \cdot), u \to \infty.$$

, Y is either a Pareto process or $\Pr(r(Y) = 0) = 1$. Since the distribution of Y and $Y - r(Y)$ uniquely determines the distribution of Y, each of the statements define the same Pareto process Y. Now, we need to verify the conditons in the first statement. The condition $\Pr(X - u \in \{x \in \mathcal{C}_0 : r(x) \geq 0\}|r(X) > u) \to \Pr(r(Y) \geq 0), u \to \infty$. Therefore, $\Pr(r(Y) \geq 0) = 1$. Assume, $\Pr(r(Y) = 0) < 1$, then, we have $\Pr(r(Y) > 0) > 0$. Let $u_1, u_2 \geq 0$, then

$$\begin{aligned}
&\Pr(X - u_1 - u_2 \in A_{v,B}, r(X) > u_2 + u_1 | r(X) > u_1) \\
=&\Pr(X - u_1 - u_2 \in A_{v,B} | r(X) > u_1 + u_2) \Pr(r(X) > u_1 + u_2 | r(X) > u_1) \\
\Rightarrow&\Pr(Y - u_2 \in A_{v,B}, r(Y) > u_2) = \Pr(Y \in A_{v,B}) \Pr(r(Y) > u_2)
\end{aligned}$$

Since the set $A_{v,B}$ with $\Pr(r(Y) = v) = \Pr(Y - r(Y) \in \partial B) = 0$ forms a $\pi-$system, we have $\Pr(Y - u2 \in \cdot | r(Y) > u_2) = \Pr(Y \in \cdot)$. Thus, Y is a Pareto process defined in (2). □

B Simulation Study and Applications

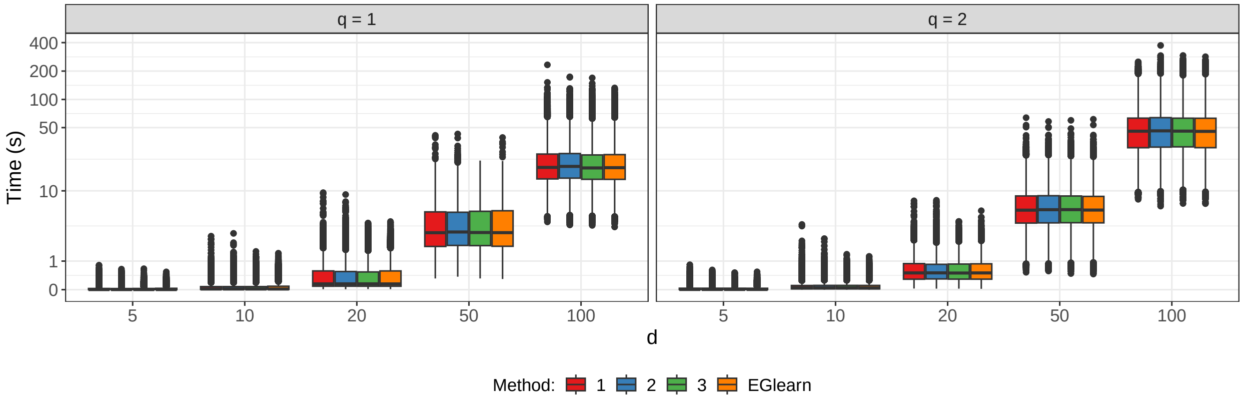

Figure 7: Boxplots of computational time in seconds for each cases and methods as in Figure 2.