Journal of Machine Learning Research 26 (2025) 1-68 Submitted 4/24; Revised 12/24; Published 2/25

Extremal graphical modeling with latent variables via convex optimization

Sebastian Engelke SEBASTIAN.ENGELKE@UNIGE.CH
Research Center for Statistics, University of Geneva

Armeen Taeb ATAEB@UW.EDU
Department of Statistics, University of Washington

Editor: Qiang Liu

Abstract

Extremal graphical models encode the conditional independence structure of multivariate extremes and provide a powerful tool for quantifying the risk of rare events. Prior work on learning these graphs from data has focused on the setting where all relevant variables are observed. For the popular class of Hüsler–Reiss models, we propose the `eglatent` method, a tractable convex program for learning extremal graphical models in the presence of latent variables. Our approach decomposes the Hüsler–Reiss precision matrix into a sparse component encoding the graphical structure among the observed variables after conditioning on the latent variables, and a low-rank component encoding the effect of a few latent variables on the observed variables. We provide finite-sample guarantees of `eglatent` and show that it consistently recovers the conditional graph as well as the number of latent variables. We highlight the improved performances of our approach on synthetic and real data.

Keywords: conditional independence, extreme value theory, latent variable model, multivariate Pareto distribution, sparsity

1. Introduction

Floods, heat waves, and financial crashes illustrate the environmental and economic hazards primarily influenced by rare, yet significant, events. Such catastrophic scenarios often result from the simultaneous occurrence of extreme values across multiple variables (Zhou, 2009; Asadi et al., 2015; Zscheischler and Seneviratne., 2017). To effectively measure and mitigate these disasters, it is essential to understand the dependencies between the various risk factors. From a mathematical perspective, this requires examining the tail dependence between the components of the random vector $X = (X_1, \ldots, X_d)$. Extreme value theory provides the theoretical foundation for extrapolations to the distributional tail of the random vector X. Within the multivariate setting, there are two different yet closely related approaches for modeling extremal data. The first method considers component-wise maxima of independent copies of X and leads to the notion of max-stable distributions (de Haan and Resnick, 1977). The second method relies on multivariate Pareto distributions that describe the random

vector X conditioned on the event that there is an extreme in one of the coordinates of X (Rootzén and Tajvidi, 2006).

Given the increasing complexity and dimensionality of contemporary data sets, identifying sparse representations for distributions of extreme events is critical for accurate modeling and risk assessment (Engelke and Ivanovs, 2021). Graphical models serve as powerful tools in achieving such sparse representations, offering clear and interpretable models for understanding dependencies among variables (Lauritzen, 1996). However, in the framework of max-stable distributions, Papastathopoulos and Strokorb (2016) highlighted limitations in developing non-trivial graphical models for their densities. On the other hand, multivariate Pareto distributions do not face these limitations. Indeed, Engelke and Hitz (2020) introduced extremal graphical models that factorize according to multivariate Pareto distributions and encode extremal conditional independence relationships, and Segers (2020) showed that extremal trees naturally arise as limits of Markov trees. For the popular Hüsler–Reiss family (Hüsler and Reiss, 1989), Hentschel et al. (2022) showed that, similarly to the Gaussian case, the sparsity pattern of an extremal graphical model can be read off from a positive semi-definite precision matrix Θ with the all-ones vector in its null space. This precision matrix Θ is derived from a transformation of the variogram matrix Γ that parameterizes a Hüsler–Reiss distribution. Several recent papers have proposed methods to learn the extremal graphical structure from data (Engelke et al., 2022c; Hu et al., 2022; Engelke and Volgushev, 2022; Chang and Allen, 2023; Wan and Zhou, 2023; Lederer and Oesting, 2023).

The study and techniques for modeling extremes have so far concentrated on scenarios where all relevant variables are directly observable. However, in many real-world situations, there exist latent variables that are not observable due to prohibitive costs or other practical constraints. Mathematically, the overall system of variables is then given by $X = (X_O, X_H)$, where X_O are the observed and X_H the latent variables, with $(O, H) = \{1, 2, \ldots, d\}$. The importance of accounting for latent factors becomes apparent in the example of a single latent variable $X_H = \{X_c\}$, where the data is generated through the one-factor model

$$X_j = X_c + \varepsilon_j, \quad j \in O.$$

Here, X_c is the common (unobserved) factor influencing all observed variables, and ε_j, $j \in O$, are independent noise terms. Suppose that the exceedances of the random vector X converge in distribution to a multivariate Pareto distribution $Y = (Y_O, Y_H)$; a concrete example where this is satisfied is when X_H is standard exponential and the noise variables are normally distributed, in which case Y has a Hüsler–Reiss distribution, but many other combinations are possible (Engelke et al., 2019). The joint vector Y can be shown to be an extremal graphical model with respect to the star graph on the left-hand side of Figure 1, where the observed variables Y_O are conditionally independent given the latent variable Y_H. However, the sub-model model of Y corresponding to the observed variables, that is, the limiting multivariate Pareto distribution arising from threshold exceedances of X_O, induces, in general, the fully connected extremal graph on the right-hand side of Figure 1, where are all the variables are conditionally dependent.

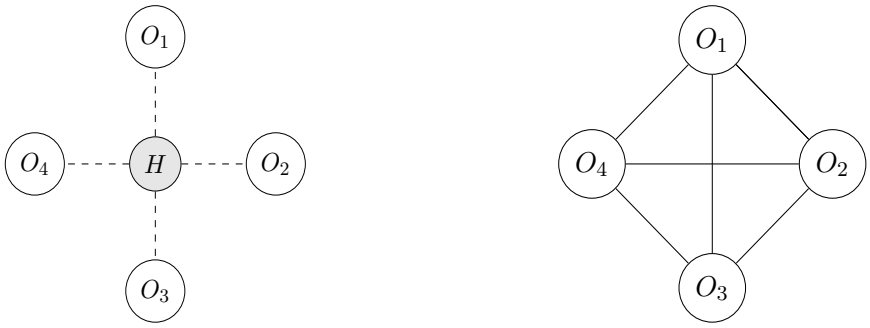

Figure 1: One-factor graph with one latent variable with four observed variables $O_1, \dots, O_4$ and one latent variable H (left) and its marginalization on the observed variables (right).

This simple example illustrates that ignoring the effect of latent variables induces confounding dependencies among the observed variables: even for a sparse joint graph of observed and latent variables, any two observed variables are dependent when conditioning on the remaining observed variables. This phenomenon also appears in many real-world applications. In such cases, a latent extremal graphical model with possibly more than one latent variable Y_H serves multiple purposes: i) it obtains the number of latent variables $h = |H|$ that summarize the effect of external phenomena on the observed variables, (ii) it identifies the residual graph structure among the observed variables after extracting away the effect of these external factors, (iii) it often yields a more sparsely represented and accurate statistical model than a model that ignores the latent variables. Latent extremal graphical models have only been studied when the graphical structure among the observed and latent variables is a tree, and where the tree structure is assumed to be known (Asenova and Segers, 2023; Röttger et al., 2023b).

1.1 Our contributions

We introduce a general latent Hüsler–Reiss graphical model where the graphical structure among the observed and latent variables as well as the number of latent variables may be arbitrary. Letting $\Theta \in \mathbb{R}^{d\times d}$ be the precision matrix, a key result that we establish is that the marginal precision matrix $\tilde{\Theta} \in \mathbb{R}^{p\times p}$ over the observed variables can be expressed in terms of blocks of Θ as

$$\tilde{\Theta} = \Theta_O - \Theta_{OH}\Theta_H^{-1}\Theta_{HO}, \quad \text{where} \quad \Theta = \begin{pmatrix} \Theta_O & \Theta_{OH} \\ \Theta_{HO} & \Theta_H \end{pmatrix}, \ \tilde{\Theta}\mathbf{1}_p = 0, \text{ and } \Theta\mathbf{1}_d = 0.$$

Here, $\mathbf{1}_r$ is the all-ones vector with r coordinates. The representation of Θ resembles the Schur complement in Gaussian latent variable graphical models (Chandrasekaran et al., 2012). However, in the Hüsler–Reiss case, the matrices Θ and $\tilde{\Theta}$ are not invertible since they have the all-one vector in their kernel, and the link between our representation and the Schur complement is therefore non-trivial.

Assuming that the conditional graph among the observed variables is sparse and that there are a few latent variables influencing the observed variables, the marginal precision matrix $\tilde{\Theta}$ is decomposed as the sum of a sparse and a low-rank matrix, i.e., $\tilde{\Theta}^\star = S^\star - L^\star$. The sparse component $S^\star := \Theta_O$ encodes the conditional graphical structure among the observed variables after conditioning on the latent variables and the low-rank component $L^\star := \Theta_{OH}\Theta_H^{-1}\Theta_{HO}$ encodes the effect of a few latent variables on the observed variables. Using this decomposition, we propose a *convex optimization procedure* named `eglatent` that provides estimates (S, L) for each term in the decomposition without knowledge of the underlying graphical structure or the number of latent variables. Compared to the latent variable graphical modeling estimator in Chandrasekaran et al. (2012), `eglatent` has the additional constraint that the matrix $S - L$ has the all-ones vector in its kernel. Due to this structural constraint that arises in extremal models, in addition to assuming that the number of latent variables is small (compared to the observed variables) and they affect many observed variables, we require new identifiability assumptions for recovering $S^\star$ and $L^\star$. Under these identifiability assumptions, we provide finite-sample consistency guarantees for our estimator, showing that our procedure recovers the conditional graph and the number of latent variables.

Figure 2 highlights the advantage of our method `eglatent` over the existing extremal graph learning method `eglearn` (Engelke et al., 2022c), which does not account for latent variables. In this synthetic example, we generated 2000 approximate observations from an extremal graphical model with $h = 2$ latent variables and a cycle graph among $p = 30$ observed variables, and fitted both methods for different values of the regularization parameters; see Section 5.1.1 for details on the setup. Compared to `eglearn`, our `eglatent` produces a better model fit on validation data and more accurate graph estimates among the observed variables in terms of F-score. Indeed, due to the latent confounding, the marginal graph among the observed variables, encoded by the zero pattern in $\tilde{\Theta}$, is dense, and thus the sparsity that `eglearn` exploits is not appropriate: the best validated `eglearn` model has 252 edges while the true graph has 30 edges. On the other hand, conditional on the latent variables, the conditional graph among the observed variables, encoded by the zero pattern in Θ_O, is sparse, and `eglatent` exploits this structure. Furthermore, `eglatent` estimates the correct number of latent variables and a near-perfect graph among the observed variables for regularization parameters with high validation likelihood. Note that in the left plot, the crosses for `eglearn` mean that the estimated graphical model is disconnected and therefore does not lead to a valid Hüsler–Reiss model. In contrast, `eglatent` always yields a valid Hüsler–Reiss model. More simulations and an application to large flight delays in the U.S. are presented in Section 5 that demonstrate the utility of our approach.

In summary, compared to the previous literature in extremal graphical modeling and Gaussian latent variable graphical modeling, our contributions are threefold. From a methodological perspective, we provide the first method to learn general extremal graphical models with latent variables. Our approach `eglatent` is based on a tractable convex optimization procedure that resembles the estimator in Chandrasekaran et al. (2012) but involves an additional constraint due to the structural properties of extremal models. From a practi-

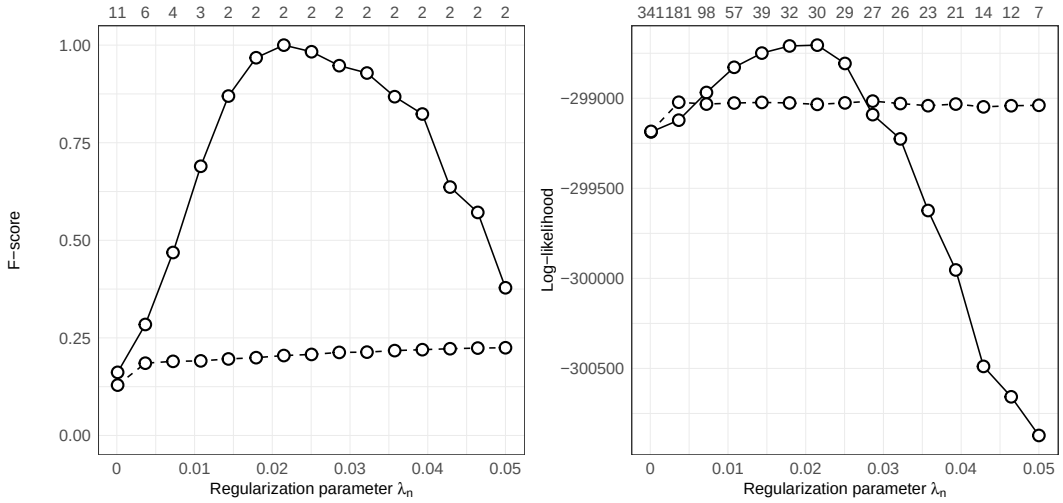

Figure 2: Left: F-score of our proposed method `eglatent` (solid line) and `eglearn` (dashed line) as a function of the regularization parameter with larger F-scores being better; top axis shows the number of estimated latent variables. Right: the likelihood of the same methods evaluated on a validation data set; the top axis shows the number of estimated edges in the latent model.

cal perspective, compared to existing extremal graphical modeling approaches that do not account for latent variables, `eglatent` often yields sparser and thus more interpretable graphical models with better fit to data. Theoretically, to arrive at our estimator `eglatent`, we prove a non-trivial Schur decomposition of the observed precision matrix of the observed variables. Further, since `eglatent` differs from the estimator in Chandrasekaran et al. (2012), we require new identifiability assumptions and conduct a more involved analysis to establish finite-sample consistency guarantees.

Our `eglatent` method is implemented in the R package `graphicalExtremes` (Engelke et al., 2022a) and all numerical results and figures can be reproduced using the code on `https://github.com/sebastian-engelke/extremal_latent_learning`.

1.2 Notation

We denote I_r as an $r \times r$ identity matrix and denote $\mathbf{1}_r$ as the all-ones vector with r coordinates. The collection of $r \times r$ symmetric matrices is denoted by $\mathbb{S}^r$. The following matrix norms are employed throughout this paper: $\|M\|_2$ denotes the spectral norm, or the largest singular value of M; $\|M\|_\infty$ denotes the largest entry in the magnitude of M; $\|M\|_\star$ denotes the nuclear norm, or the sum of the singular values of M (this reduces to the trace for positive semidefinite matrices); and $\|M\|_1$ denotes the sum of the absolute values of the entries of M. Finally, we will denote $\sigma_{\min}(M)$ as the largest non-zero singular value of M.

2. Background

2.1 Multivariate extreme value theory

Multivariate extreme value theory studies asymptotically motivated models for the largest observations of a random vector $X = (X_j : j \in V)$ with index set $V = \{1, \ldots, d\}$. Since we concentrate on models for the extremal dependence structure, we assume that the marginal distributions of X have been standardized to standard exponential distributions. In practice, this standardization can be achieved by using the marginal empirical distribution functions; see Section 3.3.1.

A multivariate Pareto distribution models the multivariate tail of the distribution of X. It is defined as the limit in the distribution of the conditional exceedances over a high threshold u, that is,

$$Y = \lim_{u \to \infty} (X - u \mid \max(X_1, \ldots, X_d) > u), \tag{1}$$

if the limit exists (Rootzén and Tajvidi, 2006). Here the simple normalization by subtracting u in each component of X is due to the exponential marginals. The random vector X is said to be in the domain of attraction of the multivariate Pareto distribution Y, which is supported on the space $\mathcal{L} = \{y \in \mathbb{R}^d : \max(y_1, \ldots, y_d) > 0\}$. Multivariate Pareto distributions are the only possible limits of threshold exceedances (Rootzén et al., 2018) and therefore a canonical model for extremes. If the convergence in (1) holds, it is easy to see that for any non-empty subset $I \subset V$, the sub-vector $X_I = (X_j : j \in I)$ is itself in the domain of attraction of a $|I|$-dimensional Pareto distribution, which we call the Ith sub-model of Y.

We now introduce the Hüsler–Reiss model, which is the most popular parametric sub-class of multivariate Pareto distributions. It can be seen as the analog of Gaussian distributions in multivariate extreme value theory, a fact, that will become apparent when studying extremal graphical models in the next section.

Definition 1 *A multivariate Pareto distribution $Y = (Y_1, \ldots, Y_d)$ is called a Hüsler–Reiss distribution parameterized by the variogram matrix Γ in the space of conditionally negative definite matrices*

$$\mathcal{C}^d = \{\Gamma \in [0, \infty)^{d \times d} : \Gamma = \Gamma^\top, \ \operatorname{diag}(\Gamma) = \mathbf{0}, \ \mathrm{v}^\top \Gamma \mathrm{v} < 0 \, \forall \, \mathbf{0} \neq \mathrm{v} \perp \mathbf{1}\}, \tag{2}$$

if its density has the form

$$f(y; \Gamma) = c_\Gamma \exp\left\{ -\frac{1}{2}(y - \mu_\Gamma)^\top \Theta (y - \mu_\Gamma) - \frac{1}{d} \sum_{i=1}^{d} y_i \right\}, \quad y \in \mathcal{L}, \tag{3}$$

where $c_\Gamma > 0$ is a normalizing constant, $\mu_\Gamma = \Pi(-\Gamma/2)\mathbf{1}_d$, and $\Pi = I_d - \mathbf{1}_d \mathbf{1}_d^\top / d$ is the projection matrix onto the orthogonal complement of the all-ones vector in d-dimensions. The matrix $\Theta = (\Pi(-\Gamma/2)\Pi)^+$ is the positive semi-definite Hüsler–Reiss precision matrix (Hentschel et al., 2022), where A^+ is the Moore–Penrose pseudoinverse of a matrix A.

The Hüsler–Reiss distribution is stable under marginalization, in the sense that for $I \subset V$, the Hüsler–Reiss sub-model corresponding to the Ith marginal is again Hüsler–Reiss distributed with parameter matrix Γ_I. While the density in (3) resembles the density of a multivariate normal distribution, we note that there are important differences. First, this function would not have finite integral on $\mathbb{R}^d$ because of the second term in the exponential, and the restriction to the subset $\mathcal{L}$ is crucial. Second, the precision matrix Θ is of rank $d-1$, which complicates theoretical and practical considerations.

An important summary statistic of the dependence structure in multivariate Pareto distributions is the extremal variogram (Engelke and Volgushev, 2022). It takes a similar role as the covariance matrix in the non-extremal world.

Definition 2 *For a multivariate Pareto distribution $Y = (Y_j : j \in V)$ the extremal variogram rooted at node $m \in V$ is defined as the matrix $\Gamma^{(m)}$ with entries*

$$\Gamma^{(m)}_{ij} = \mathrm{Var}\left\{Y_i - Y_j \mid Y_m > 1\right\}, \quad i, j \in V,$$

whenever the right-hand side is finite.

If Y follows a Hüsler–Reiss distribution with parameter matrix Γ, it can be checked that the extremal variogram matrices coincide for all $m \in V$, and that they satisfy

$$\Gamma = \Gamma^{(1)} = \cdots = \Gamma^{(d)}. \tag{4}$$

We use this fact later to combine empirical estimators of the extremal variograms rooted at the different nodes to obtain a more efficient joint estimator of Γ.

2.2 Extremal graphical models

Conditional independence for multivariate Pareto distributions Y is non-standard since it is defined on the space $\mathcal{L}$, which is not a product space. Engelke and Hitz (2020) therefore define a new notion of extremal conditional independence using the auxiliary vectors $Y^{(m)}$, for $m \in \{1, \ldots, d\}$, defined as Y conditioned on the event that $\{Y_m > 0\}$. For non-empty subsets $A, B, C \subset V$, we say that Y_A is conditionally independent of Y_B given Y_C, denoted by $Y_A \perp_e Y_B \mid Y_C$, if for all auxiliary random vectors, we have the corresponding statement in the usual sense, that is,

$$Y^{(m)}_A \perp\!\!\!\perp Y^{(m)}_B \mid Y^{(m)}_C \quad \text{for all } m \in V.$$

It can be shown that requiring the relation above is equivalent to requiring the existence of a single $m \in V$ for which $Y^{(m)}_A \perp\!\!\!\perp Y^{(m)}_B \mid Y^{(m)}_C$ (Engelke et al., 2022b).

Let $\mathcal{G} = (V, E)$ be an undirected graph with nodes $V = \{1, \ldots, d\}$ and edge set $E \subset V \times V$. Using the new notion of conditional independence, an extremal graphical model on $\mathcal{G}$ is a multivariate Pareto distribution Y that satisfies the extremal pairwise Markov property on $\mathcal{G}$, that is,

$$Y_i \perp_e Y_j \mid Y_{V \setminus \{i,j\}} \quad \text{if } (i, j) \notin E.$$

Engelke and Hitz (2020) show that this definition is natural in the sense that it enables a Hammersley–Clifford theorem showing that densities factorize into lower-dimensional terms on the cliques of the graph.

For a multivariate Gaussian distribution with covariance matrix Σ, the conditional dependence relationships, or equivalently the edges in the Gaussian graphical model can be identified from the nonzeros of the precision matrix Σ^{-1}. A similar property holds for extremal graphical models if Y follows a Hüsler–Reiss distribution, where the matrix Θ in Definition 1 plays a key role.

Proposition 3 (Lemma 1 and Proposition 3 of Engelke and Hitz (2020)) *Let $Y \in \mathbb{R}^d$ follow a Hüsler–Reiss distribution with precision matrix Θ. Then,*

$$Y_i \perp_e Y_j \mid Y_{V\setminus\{i,j\}} \Leftrightarrow \Theta_{ij} = 0. \tag{5}$$

A consequence of Proposition 3 is that for a Hüsler–Reiss graphical model on an arbitrary connected graph $\mathcal{G}$, we can read off the graph structure from the zero pattern of the precision matrix Θ.

Finally, we note that an important property of an extremal graphical model is that if Y possesses a density that factorizes on the graph $\mathcal{G}$, then $\mathcal{G}$ must *necessarily be connected* (Engelke and Hitz, 2020). The state-of-the-art structure learning methods for extremal data (Engelke et al., 2022c; Wan and Zhou, 2023) can yield disconnected graphs that thus do not always yield a valid distribution (see the example in Figure 1). For a detailed review of recent progress on extremal graphical models, we refer to Engelke et al. (2024a). In the next section, we present our approach for structure learning, which can handle latent variables and always yields a valid distribution.

3. Latent Hüsler–Reiss models and the `eglatent` method

3.1 Latent Hüsler–Reiss models

In the illustrative example in the introduction, we presented a Hüsler–Reiss model with a single latent variable and a very simple graphical structure among the observed variables. We next introduce a latent Hüsler–Reiss model with a general extremal graphical structure and any number of latent variables. In what follows, let $X_O \in \mathbb{R}^p$ be the collection of observed variables, $X_H \in \mathbb{R}^h$ be a collection of latent variables, and put $d := p + h$.

Definition 4 (Latent Hüsler–Reiss models) *Suppose that the random vector $X = (X_O, X_H) \in \mathbb{R}^d$, indexed by $V = (O, H)$, is in the domain of attraction of a Hüsler–Reiss distribution $Y \in \mathbb{R}^d$ in the sense of (1) with variogram and precision matrices, and corresponding extremal graphical structure*

$$\Gamma = \begin{pmatrix} \Gamma_O & \Gamma_{OH} \\ \Gamma_{HO} & \Gamma_H \end{pmatrix}, \qquad \Theta = \begin{pmatrix} \Theta_O & \Theta_{OH} \\ \Theta_{HO} & \Theta_H \end{pmatrix}, \quad \text{and} \quad \mathcal{G} = (V, E),$$

respectively. Here $\Theta = (\Pi(-\Gamma/2)\Pi)^+$, Γ_O and Θ_O are $p \times p$-dimensional symmetric matrices, and $E = \{(i,j) : i,j \in V, i \neq j, \Theta_{ij} \neq 0\}$. We then say that Y is a latent Hüsler–Reiss model,

and we note that the observed variables X_O are in the domain of attraction of a Hüsler–Reiss model with variogram Γ_O.

Note that Θ_O and Θ_H in the above definition are positive definite since $\Gamma \in \mathcal{C}^d$; see Definition 1 and Engelke and Hitz (2020, Appendix B). Latent Hüsler–Reiss models have been studied only for very simple graphs, namely tree structures (Asenova et al., 2021; Röttger et al., 2023b) and block graphs (Asenova and Segers, 2023). All of the above methods assume the underlying graph structure among the observed and latent variables and the number of latent variables to be known, which is rarely realistic in practice. To handle more general graphs, we establish the following theorem, which relates the marginal distribution of the observed variables to components of the precision matrix Θ.

Theorem 5 *Let $\tilde{\Pi} = I_p - \mathbf{1}_p\mathbf{1}_p^T/p$ be the projection matrix onto the orthogonal complement of the all-ones vector in p dimensions. Then, the precision matrix $\tilde{\Theta} \in \mathbb{R}^{p\times p}$ of the observed variables of a latent Hüsler–Reiss model with variogram matrix Γ satisfies*

$$\tilde{\Theta} = (\tilde{\Pi}(-\Gamma_O/2)\tilde{\Pi})^+ = \Theta_O - \Theta_{OH}\Theta_H^{-1}\Theta_{HO}. \tag{6}$$

While it is not possible to observe the joint precision matrix Θ or any of its components directly, Theorem 5 provides a useful decomposition of the observable precision matrix $\tilde{\Theta}$ into the difference of two terms, each term involving the components of Θ. By the property in (5), we have for any $i, j \in O$ that

$$Y_i \perp_e Y_j \mid Y_H, Y_{O\setminus\{i,j\}} \quad \Leftrightarrow \quad [\Theta_O]_{i,j} = 0.$$

Thus, the first term Θ_O in decomposition (6) specifies the conditional independencies among the observed variables after conditioning on the latent variables. Moreover, the sparsity pattern of Θ_O encodes the residual graph $\mathcal{G}_O$ among the observed variables after extracting the influence of the latent variables. Here, $\mathcal{G}_O = (O, E_O)$ is a subgraph of $\mathcal{G}$ restricted to the observed variables where $E_O = \{(i,j) \in E, i, j \in O\}$. The second term $\Theta_{OH}\Theta_H^{-1}\Theta_{HO}$ in decomposition (6) serves as a summary of the marginalization of the latent variables Y_H and encodes their effect on the observed variables. The rank of this matrix is equal to the number of latent variables. The overall term $\Theta_O - \Theta_{OH}\Theta_H^{-1}\Theta_{HO}$ is a Schur complement with respect to Θ_H.

As an illustration, consider the extremal graph on the left-hand side of Figure 1. Here, the matrix Θ_O is diagonal. Furthermore, the matrix $\Theta_{OH}\Theta_H^{-1}\Theta_{HO}$ has rank equal to one with all of its entries being nonzero. Note that $\tilde{\Theta}$ generally consists of all nonzero entries and hence the marginal graphical structure among the observed variables on the right-hand side of Figure 1 is fully connected.

3.2 Sparse plus low-rank decomposition

In this paper, we consider a latent Hüsler–Reiss graphical model where the subgraph $\mathcal{G}_O$ among the observed variables is sparse and the number of latent variables is small relative to the number of observed variables, that is, $h \ll p$. This modeling assumption is often natural in real-world applications. For example, Chandrasekaran et al. (2012) and Taeb and Chandrasekaran (2016) showed that a large fraction of the conditional dependencies among stock returns can be explained by a small number of latent variables and interpreted these to be correlated to exchange rate and government expenditures. In a similar spirit, Taeb et al. (2017) demonstrated that the California reservoir network is sparsely connected after accounting for a few latent factors, and interpreted these latent factors to be highly correlated to environmental variables such as drought level and precipitation.

In the case of extremes, a sparse subgraph $\mathcal{G}_0$ and the presence of only a few latent variables in the model translate to a latent Hüsler–Reiss model with matrix Θ_O being sparse, the matrix $\Theta_{OH}\Theta_H^{-1}\Theta_{HO}$ being low-rank, and thus the observed precision matrix $\tilde{\Theta}$ being decomposed as a sparse plus low-rank matrix having zero row sums. Notice that the matrix $\tilde{\Theta}$ will generally be dense due to the additional low-rank term $\Theta_{OH}\Theta_H^{-1}\Theta_{HO}$, highlighting how the latent variables induce many confounding dependencies among the observed variables (see Figure 1), and how structure learning procedures that impose sparsity on the precision matrix $\tilde{\Theta}$ will generally not perform well.

In summary, we can cast the problem of learning a latent Hüsler–Reiss graphical model as obtaining a sparse plus low-rank decomposition of the precision matrix $\tilde{\Theta}$ of the observed variables. The sparse component provides the residual graphical structure of the observed variables after accounting for the latent variables, the rank of the low-rank component provides the number of latent variables, and the overall sum provides a compact model of the observed variables that can be used for downstream tasks. In the following section, we propose a convex optimization procedure to accurately estimate each of these components from data.

Finally, we note that in the setting where the observed and latent variables are jointly Gaussian, Chandrasekaran et al. (2012) also models the precision matrix among the observed variables as a sum of a sparse and a low-rank matrix. Analogous to our setting, the sparse component encodes the subgraph of the observed variables and the low-rank component encodes the effect of the latent variables on the observed variables. An important distinguishing feature with our extremal setting however is that in the Gaussian context, the resulting sum is not constrained to have zero row sum. As we describe in Section 3.3, the additional subspace constraint in our extremal setting results in a different estimation procedure and assumptions for statistical consistency.

3.3 Inference for latent Hüsler–Reiss graphical models

Let $X = (X_O, X_H)$ be a collection of observed and latent variables in the domain of attraction of a latent Hüsler–Reiss graphical model with a sparse subgraph among the observed variables and a small number of latent variables; we will specify the sparsity level and the number of

latent variables in our theoretical results. Let $\Gamma^\star$ be the underlying population variogram matrix and $\Theta^\star$ be the population precision matrix with components $\Theta^\star_O, \Theta^\star_{OH}$ and $\Theta^\star_H$. Let $\tilde{\Theta}^\star$ be the precision matrix among the observed variables. From Theorem 9, we have that $\tilde{\Theta}^\star = S^\star - L^\star$ where $S^\star := \Theta^\star_O$ is a sparse matrix and $L^\star := \Theta^\star_{OH}{\Theta^\star_H}^{-1}\Theta^\star_{HO}$ is a low-rank matrix. Here, the support of $S^\star$ encodes the subgraph among the observed variables and the rank of $L^\star$ encodes the number of latent variables. We will propose a convex optimization procedure to estimate the matrices $(S^\star, L^\star)$ from data.

3.3.1 Empirical extremal variogram matrix

An important ingredient of our procedure is an empirical estimate for the extremal variogram matrix $\Gamma^\star_O$. To arrive at our estimate, define for any $m \in O$, the population extremal variogram matrix $\Gamma^{\star,(m)}_O$ rooted at the node m; see Definition 2. Suppose we have n independent and identically distributed samples $\{X^{(t)}_O\}_{t=1}^n \subseteq \mathbb{R}^p$ of the observed variables X_O. Then, a natural estimate $\hat{\Gamma}^{(m)}_O$ for $\Gamma^{\star,(m)}_O$ is given by

$$\hat{\Gamma}^{(m)}_{ij} := \widehat{\mathrm{Var}}\Big(\log(1 - \hat{F}_i(X^{(t)}_i)) - \log(1 - \hat{F}_j(X^{(t)}_j)) : \hat{F}_m(X^{(t)}_m) \geq 1 - k/n \Big), \quad i, j \in O.$$

Here, $\widehat{\mathrm{Var}}$ denotes the sample variance, and k is the number of extreme samples considered in the conditioning event, which can be viewed as the *effective sample size*. Since in Section 2 we assumed that X has standard exponential margins, for $i \in O$, $t \in \{1, \dots, n\}$, inside the variance we normalize the i-th entry of the t-th observation empirically by $-\log(1-\hat{F}_i(X^{(t)}_i))$, where $\hat{F}_i$ denotes the empirical distribution function of $X^{(1)}_i, \dots, X^{(n)}_i$. As (4) establishes that the empirical variogram matrix rooted at node m coincides with the true variogram matrix $\Gamma^\star_O$ for every m, a natural empirical estimator of this matrix is

$$\hat{\Gamma}_O := \frac{1}{p} \sum_{m=1}^{p} \hat{\Gamma}^{(m)}_O. \tag{7}$$

Under the assumption that $k \to \infty$ and $k/n \to 0$, and mild conditions on the underlying data generation, this estimator can be shown to be consistent for $\Gamma^\star_O$ (Engelke and Volgushev, 2022). Moreover, Engelke et al. (2022c) derive finite sample concentration bounds for $\hat{\Gamma}_O$ that can be used for high-dimensional consistency results. We refer to Appendix G for details on the assumptions and results.

3.3.2 Parameter estimation and structure learning

For structure learning in Hüsler–Reiss models, formulating optimization problems in the precision domain leads to computationally efficient procedures. Indeed, the precision matrix estimate obtained from plugging in the empirical extremal variogram $\hat{\Gamma}_O$ in place of $\Gamma^\star_O$ in

the expression $\tilde{\Theta}^\star = (\tilde{\Pi}(-\Gamma^\star_O/2)\tilde{\Pi})^+$ is the minimizer of the convex problem

$$\begin{aligned} \hat{\Theta} = \underset{\Theta \in \mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det\left(U^T\Theta U\right) - \frac{1}{2}\operatorname{tr}(\Theta\hat{\Gamma}_O), \\ \text{s.t.} \quad & \Theta \succeq 0 \quad , \quad \Theta\mathbf{1}_p = 0, \end{aligned} \tag{8}$$

where the matrix $U \in \mathbb{R}^{p\times(p-1)}$ consists of the first $p-1$ left singular vectors of $\tilde{\Pi}$ so that $UU^T = \tilde{\Pi}$; see Appendix C for a formal proof. The constraint $\succeq 0$ imposes positive semi-definiteness, $\mathbb{S}^p$ denotes the space of symmetric $p \times p$ matrices, and the constraint $\Theta\mathbf{1}_p = 0$ ensures that Θ has zero row sum. The above optimization problem corresponds to the surrogate maximum likelihood estimator of the Hüsler–Reiss distribution; for more details on this justification we refer to Röttger et al. (2023b).

The formulation in terms of the precision matrix Θ opens the door to various regularized estimation methods. Röttger et al. (2023b) solve (8) under the additional constraint that $\Theta_{ij} \leq 0$ for all $i, j \in V$ to ensure a from of positive dependence. For a graph $\mathcal{G} = (V, E)$, in order to obtain a graph structured estimate of Γ, Hentschel et al. (2022) solve a matrix completion problem that corresponds to (8) under the constraint $\Theta_{ij} = 0$ for $(i, j) \notin E$. In the context of structure learning without latent variables, Engelke et al. (2022c) and Wan and Zhou (2023) add an ℓ_1 penalty to the loss function akin to the graphical Lasso.

In the setting with latent variables, we rely on the sparse plus low-rank decomposition described in Section 3.1. We, therefore, search over the space of precision matrices Θ that can be decomposed as $\Theta = S - L$ to identify a sparse matrix S and a low-rank matrix L, whose difference has zero row sum and yields a small surrogate negative likelihood. Motivated by the estimator for Gaussian latent variable graphical modeling (Chandrasekaran et al., 2012), we introduce the `eglatent` method that solves the following regularized *convex* likelihood problem for some $\lambda_n, \gamma \geq 0$:

$$\begin{aligned} (\hat{S}, \hat{L}) = \underset{S\in\mathbb{S}^p, L\in\mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det(U^T(S-L)U) - \operatorname{tr}((S-L)\hat{\Gamma}_O/2) + \lambda_n(\|S\|_1 + \gamma\operatorname{tr}(L)), \\ \text{s.t.} \quad & S - L \succeq 0, L \succeq 0, (S-L)\mathbf{1}_p = 0. \end{aligned} \tag{9}$$

Here, $\hat{S}$ and $\hat{L}$ are estimates for the population quantities $S^\star$ and $L^\star$, respectively. The matrix $\hat{S} - \hat{L}$ represents an estimated precision matrix among the observed variables. By the constraints in (9) and the property of logdet functions, $\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ is the null space of $\hat{S} - \hat{L}$ and $\hat{S} - \hat{L}$ always specifies a valid Hüsler–Reiss model.

The function $\|\cdot\|_1$ denotes the ℓ_1 norm that promotes sparsity in the matrix S (Friedman et al., 2007). The role of the trace penalty on L is to promote low-rank structure (Fazel et al., 2004). The regularization parameter γ provides a trade-off between the graphical model component and the latent component. In particular, for very large values of γ, `eglatent` produces $\hat{L} = 0$ so that no latent variables are included in the model. As γ decreases, the number of latent variables increases and correspondingly the number of edges in the residual graphical structure decreases. The regularization parameter λ_n provides overall control of the trade-off between the fidelity of the model to the data and the complexity of the model,

and thus naturally depends on the sample size. For $\lambda_n, \gamma \geq 0$, `eglatent` is a convex program that can be solved efficiently.

While our `eglatent` estimator (9) resembles the one in Chandrasekaran et al. (2012), there is an important distinguishing feature. Specifically, in contrast to the estimator in Chandrasekaran et al. (2012), our estimator imposes the constraint $(S - L)\mathbf{1}_p = 0$ so that the resulting model is a valid Hüsler–Reiss model. As a result of this additional constraint, the log-determinant term in our objective is also different in that it projects $S - L$ onto the space of matrices that have zero row/column sum. Since our estimator is different, we need additional assumptions and more involved analysis to establish consistency guarantees; see Section 4 for more details.

Remark 6 *A challenge with the optimization in* (8)*, both theoretically and numerically, is the fact that the matrices range in the space of positive semi-definite matrices with zero row sum. This factor indeed seems to prohibit direct structure learning without latent variables (i.e., setting* $L = 0$ *in* (9) *to obtain a graphical lasso analog) where the estimated graphical structure can be rather different than the true graphical structure; see the discussion in Engelke et al. (2022c, Section 7). To circumvent this issue, Engelke et al. (2022c) and Wan and Zhou (2023) solve slightly different problems to obtain accurate graph estimation, although their estimated graphs do not always yield valid Hüsler–Reiss models. Remarkably, the addition of the low-rank component* L *in the* `eglatent` *estimator* (9) *solves these issues. Indeed, we will show that* `eglatent` *consistently recovers the subgraph among the observed variables and the number of latent variables, and matches the performance of existing procedures (Engelke et al., 2022c; Wan and Zhou, 2023) for learning an accurate model when no latent variables are present.*

4. Consistency guarantees for `eglatent`

Recall from Section 3.3 that we denote by $S^\star$ the population matrix encoding the graphical structure among the observed variables conditioned on the latent variables, and by $L^\star$ the population matrix encoding the effect of a few latent variables on the observed variables. Further, $\tilde{\Theta}^\star = S^\star - L^\star$ represents the marginal precision matrix in the Hüsler–Reiss model over the observed variables. In this section, we state a theorem to prove that the estimates of `eglatent` in (9) provide, with high probability, the correct graphical structure among the observed variables, the correct number of latent variables, and an accurate extremal model. Stated mathematically, we show with high probability that (i) the sign-pattern of $\hat{S}$ is the same as that of $S^\star$, i.e., $\operatorname{sign}(\hat{S}) = \operatorname{sign}(S^\star)$, where $\operatorname{sign}(0) = 0$; (ii) the rank of $\hat{L}$ is the same as that of $L^\star$, i.e., $\operatorname{rank}(\hat{L}) = \operatorname{rank}(L^\star)$; and (iii) the estimated precision model $\hat{S} - \hat{L}$ closely approximates the true precision matrix $\tilde{\Theta}^\star$, i.e., $\hat{S} - \hat{L} \approx \tilde{\Theta}^\star$. Our analysis requires assumptions on the population model so that the matrices $S^\star$ and $L^\star$ are identifiable from their sum, and that the number of effective samples k is of order $k \gtrsim p^2 \log(p)$.

4.1 Technical setup

As `eglatent` is solved in the precision matrix parameterization, the conditions for our theorems are naturally stated in terms of the precision matrix $S^\star - L^\star$. The assumptions are similar in spirit to convex relaxation methods for Gaussian latent-variable graphical model selection (Chandrasekaran et al., 2012), although some conditions are new due to the zero row and column sum structure of the observed precision matrix $S^\star - L^\star$.

To ensure correct graph recovery and correct number of latent variables, we seek an estimate $(\hat{S}, \hat{L})$ from `eglatent` such that $\text{support}(\hat{S}) = \text{support}(S^\star)$ and $\text{rank}(\hat{L}) = \text{rank}(L^\star)$. Building on both classical statistical estimation theory, as well as the recent literature on high-dimensional statistical inference, a natural set of conditions for accurate parameter estimation, is to assume that the curvature of $S^\star - L^\star$ is bounded in certain directions. The curvature is governed by the modified Hessian of the surrogate log-likelihood loss at $S^\star - L^\star$:

$$\mathbb{I}^\star := \left(S^\star - L^\star + \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top\right)^{-1} \otimes \left(S^\star - L^\star + \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top\right)^{-1},$$

where $\otimes$ denotes a Kronecker product between matrices, and $\mathbb{I}^\star$ may be viewed as a map from $\mathbb{S}^p$ to $\mathbb{S}^p$. The matrix $\mathbb{I}^\star$ modifies the Hessian of the surrogate log-likelihood loss $(S^\star - L^\star)^+ \otimes (S^\star - L^\star)^+$, where the addition of the term $\frac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top$ (a dual parameter of the program (9)) helps to compactify the assumptions we place in our population model.

We impose conditions so that $\mathbb{I}^\star$ is well-behaved when applied to matrices of the form $S - S^\star - (L - L^\star + t\mathbf{1}_p\mathbf{1}_p^\top)$. Here, S is in the neighborhood of $S^\star$ restricted to sparse matrices, L is in the neighborhood of $L^\star$ restricted to low-rank matrices, and $t\mathbf{1}_p\mathbf{1}_p^\top$ is a dual parameter for some $t \in \mathbb{R}$ due to the constraint $(S-L)\mathbf{1}_p = 0$ that appears in the analysis of (9). These local properties of $\mathbb{I}^\star$ around $S^\star - (L^\star + t\mathbf{1}_p\mathbf{1}_p^\top)$ are conveniently stated in terms of tangent spaces to algebraic varieties of sparse and low-rank matrices. In particular, the tangent space of a matrix M with r non-zero entries with respect to the algebraic variety of $p \times p$ matrices with at most r non-zeros is given by

$$\Omega(M) := \{N \in \mathbb{R}^{p\times p} : \text{support}(N) \subseteq \text{support}(M)\}.$$

Moreover, the tangent space at a rank-r matrix M with respect to the algebraic variety of $p \times p$ matrices with rank less than or equal to r is given by:

$$T(M) := \{N_R + N_C : N_R, N_C \in \mathbb{R}^{p\times p}, \\ \text{row-space}(N_R) \subseteq \text{row-space}(M), \text{col-space}(N_C) \subseteq \text{col-space}(M)\}.$$

For more discussion on the tangent spaces of sparse and low-rank matrices, see Chandrasekaran et al. (2012). In the next section, we describe conditions on the population Hessian $\mathbb{I}^\star$ in terms of tangent spaces $\Omega(S^\star)$ and $T(L^\star)$. Under these conditions, we present a theorem in Section 4.4 showing that the convex program provides accurate estimates. For notational simplicity, we let $\Omega^\star := \Omega(S^\star)$ and $T^\star := T(L^\star)$. Finally, the linear operators $\mathcal{A} : \mathbb{S}^p \times \mathbb{S}^p \to \mathbb{S}^p$ and its adjoint $\mathcal{A}^\dagger : \mathbb{S}^p \to \mathbb{S}^p \times \mathbb{S}^p$ are defined as:

$$\mathcal{A}(M, N) := (M - N), \quad \mathcal{A}^\dagger(Q) := (Q, Q). \tag{10}$$

4.2 Conditions on the Hessian $\mathbb{I}^\star$

Given a norm $\|\cdot\|_\Psi$ on $\mathbb{S}^p \times \mathbb{S}^p$, we first consider a classical condition in statistical estimation literature, which is to control the minimum gain of the Hessian $\mathbb{I}^\star$ restricted to a subspace $\mathbb{Q} \subseteq \mathbb{S}^p \times \mathbb{S}^p$ as follows:

$$\chi(\mathbb{Q}, \|\cdot\|_\Psi) := \min_{\substack{Z \in \mathbb{H} \\ \|Z\|_\Psi = 1}} \|\mathcal{P}_\mathbb{Q} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_\mathbb{Q}(Z)\|_\Psi, \tag{11}$$

where $\mathcal{P}_\mathbb{Q}$ denotes the projection operator onto the subspace $\mathbb{Q}$ and the linear maps $\mathcal{A}$ and $\mathcal{A}^\dagger$ are defined in (10). The quantity $\chi(\mathbb{Q}, \|\cdot\|_\Psi)$ insures that the Hessian is well-conditioned restricted to the image $\mathcal{A}\mathbb{Q}$. The remaining condition we impose on $\mathbb{I}^\star$ are in the spirit of irrepresentability-type conditions that are frequently employed in high-dimensional estimation problems (Meinshausen and Buhlmann, 2006; Wainwright, 2009; Zhao and Yu, 2006; Ravikumar et al., 2008; Candès and Recht, 2012; Chandrasekaran et al., 2012). Specifically, we control the inner-product between elements in $\mathcal{A}\mathbb{Q}$ and $\mathcal{A}\mathbb{Q}^\perp$ as quantified by the metric induced by $\mathbb{I}^\star$ via the following quantity:

$$\varphi(\mathbb{Q}, \|\cdot\|_\Psi) := \max_{\substack{Z \in \mathbb{Q} \\ \|Z\|_\Psi = 1}} \|\mathcal{P}_{\mathbb{Q}^\perp} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_\mathbb{Q} (\mathcal{P}_\mathbb{Q} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_\mathbb{Q})^{-1}(Z)\|_\Psi. \tag{12}$$

The operator $(\mathcal{P}_\mathbb{Q} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_\mathbb{Q})^{-1}$ in (12) is well-defined if $\chi(\mathbb{Q}, \|\cdot\|_\Psi) > 0$, since this latter condition implies that $\mathbb{I}^\star$ is injective restricted to $\mathcal{A}\mathbb{Q}$. The quantity $\varphi(\mathbb{Q}, \|\cdot\|_\Psi)$ being small implies that any element of $\mathcal{A}\mathbb{Q}$ and any element of $\mathbb{Q}^\perp$ have a small inner-product (in the metric induced by $\mathbb{I}^\star$).

A natural approach to controlling the condition of the Hessian $\mathbb{I}^\star$ around $S^\star - L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$ is to bound the quantities $\chi(\mathbb{Q}^\star, \|\cdot\|_\Psi)$ and $\varphi(\mathbb{Q}^\star, \|\cdot\|_\Psi)$ for $\mathbb{Q}^\star = \Omega^\star \times (T^\star \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top))$. However, a complication that arises with tangent spaces to low-rank varieties is that they are locally smooth. To account for this curvature, we bound distances of nearby tangent spaces via the following induced norm:

$$\rho(T_1, T_2) := \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2.$$

The quantity $\rho(T_1, T_2)$ measures the sine of the largest angle between T_1 and T_2. Using this approach for bounding nearby tangent spaces, we consider subspaces $\mathbb{Q}' = \Omega^\star \times (T' \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ for all T' close to $T^\star$ as measured by ρ. For $\omega \in (0, 1)$, we bound $\chi(\mathbb{Q}', \|\cdot\|_\Psi)$ and $\varphi(\mathbb{Q}', \|\cdot\|_\Psi)$ in the sequel for all subspaces $\mathbb{Q}'$ in the following set:

$$U(\omega) = \{\Omega^\star \times (T' \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)) \mid \rho(T', T^\star) \leq \omega\}.$$

We control the quantities $\chi(\mathbb{Q}', \|\cdot\|_\Psi)$ and $\varphi(\mathbb{Q}', \|\cdot\|_\Psi)$ using the dual norm of the regularizer $\|S\|_1 + \gamma \operatorname{tr}(L^\star)$:

$$\Phi_\gamma(S, L) := \max\left\{\|S\|_1, \frac{\|L\|_2}{\gamma}\right\}.$$

As the dual norm $\max\{\|S\|_1, \frac{\|L\|_2}{\gamma}\}$ plays a central role in the optimality conditions of (9), controlling the quantities $\chi(\mathbb{Q}', \|\cdot\|_{\Phi_\gamma})$ and $\varphi(\mathbb{Q}', \|\cdot\|_{\Phi_\gamma})$ leads to a natural set of conditions that guarantee the consistency of the estimates produced by (9). In summary, given a fixed set of parameters $(\omega, \gamma) \in (0,1) \times \mathbb{R}_+$, we assume that $\mathbb{I}^\star$ satisfies the following conditions, where $F = \mathcal{P}_{T^{\star\perp}}(1/p\mathbf{1}_p\mathbf{1}_p^\top)/\|\mathcal{P}_{T^{\star\perp}}(1/p\mathbf{1}_p\mathbf{1}_p^\top)\|_2$ and $\|\mathbb{I}^\star\|_2$ denotes the spectral norm of the operator $\mathbb{I}^\star$.

Assumption 1 $\inf_{\mathbb{Q}' \in U(\omega)} \chi(\mathbb{Q}', \Phi_\gamma) \geq \alpha$ *for some* $\alpha > 8\omega \max\{\gamma, 1\}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega + 1)$.

Assumption 2 $\sup_{\mathbb{Q}' \in U(\omega)} \varphi(\mathbb{Q}', \Phi_\gamma) \leq 1 - \nu$ *for some* $\nu \in [4\omega, 1)$.

Chandrasekaran et al. (2012) impose a sufficient set of conditions, and prove that they imply conditions similar to Assumptions 1-2 (see Proposition 3.3 in Chandrasekaran et al. (2012)). A key distinction between our conditions and the implied conditions in Chandrasekaran et al. (2012) is that our subspace $\mathbb{Q}'$ also contains the directions $\text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$. This distinction arises from the additional zero row-sum constraint in our estimator which introduces the dual parameter $t\mathbf{1}_p\mathbf{1}_p^\top$. Moreover, we require the following condition for how far $\text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ deviates from $T^\star$:

Assumption 3 $\kappa^\star := \|\mathcal{P}_{T^{\star\perp}}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2 \in \left(\omega, \min\left\{4\nu, \frac{\alpha}{8\max\{\gamma,1\}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega+1)} - \omega\right\}\right)$.

Assumption 3 is also a new condition relative to Chandrasekaran et al. (2012). This assumption ensures that $k^\star$ not so small so that $L^\star$ and the dual parameter $t\mathbf{1}_p\mathbf{1}_p^\top$ can be distinguished from one another. Assumption 3 also ensures that $\kappa^\star$ is not too large. This condition comes from the optimality conditions of (9), which involve controlling the size of the inner product of elements in $\text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ and in $T^{\star\perp}$. Further, bounding $\kappa^\star$ allows the size of t to be controlled.

Remark 7 (Dependency on h and graph structure) *The dependence on the number of latent variables h and the density of the graphical structure among the observed variables conditioned on the latent variables does not appear explicitly in Assumptions 1–3, but is implicit in the quantities α, ν. Indeed, as larger h and denser graphical structures increase the dimensions of the tangent spaces $T^\star$ and $\Omega^\star$, respectively, they result in smaller α, ν. In Appendix F, we provide conditions on the Hessian $\mathbb{I}^\star$ that do not depend on γ and measure the behavior of $\mathbb{I}^\star$ restricted to individual subspaces $\Omega^\star$ and $T^\star$ (rather than their coupling as in Assumptions 1–2). With these conditions and when the latent variables affect most of the observed variables (see also the discussion in Section 4.1), we prove in Appendix F that as long as $d^\star\sqrt{h/p} = \mathcal{O}(1)$, there exists a choice of γ that satisfies Assumption 1–3. Here,*

$$d^\star := \max_i \sum_j \mathbb{I}[S^\star_{ij} \neq 0] \tag{13}$$

is the maximum degree of the graphical structure among the observed variables. For instance, for the following nontrivial classes of models the above condition holds:

- *Polynomial degree: the maximum degree $d^\star$ grows at most polynomially with p, that is, $d^\star = \mathcal{O}(p^q)$, and the number of latent variables satisfies $h = \mathcal{O}\left(p^{1-q}\right)$, where $q \in (0,1)$. Here, consistent estimation is possible even when the graph structure is complex.*

- *Bounded degree: we have $d^\star = \mathcal{O}(1)$ so that $h = \mathcal{O}(p)$. Here again, consistent estimation of the underlying graphical structure among the observed variables is possible even when the number of latent variables is in the same order as the number of observed variables.*

Remark 8 (Choice of γ) *We make two observations. First, a smaller range of values of γ naturally leads to larger α and ν. Second, intuitively, the choice of γ should decrease with a larger h so that less penalty is imposed on the rank of $\hat{L}$, and it should increase with larger $d^\star$ so that $\hat{L}$ does not contain some of the components of $S^\star$. To formalize this intuition, we consider the setting described in the previous paragraph. We show in Appendix F that the lower-bound on the range of values of γ that satisfy Assumptions 1–3 scales with $d^\star$, and the upper-bound is in the order $\sqrt{p/h}$.*

4.3 When do Assumptions 1–3 on the Hessian hold? Connections to identifiability

In this section, we provide concrete examples of latent extremal models that satisfy Assumptions 1–3 for some choices of $\alpha > 0$, $\nu \in [0,1)$, $\omega \in (0,1)$ and $\gamma > 0$. To arrive at such models, we must intuitively understand when the matrices $S^\star$, $L^\star$, and $t\mathbf{1}_p\mathbf{1}_p^\top$ are identifiable from their sum for some $t \in \mathbb{R}$ (recall, the term $t\mathbf{1}_p\mathbf{1}_p^\top$ arises from the zero row-sum constraint). Since the matrix $t\mathbf{1}_p\mathbf{1}_p^\top$ has rank equal to one and is thus also low-rank, we consider a combined term $L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$. Two identifiability issues arise: the first is to distinguish $S^\star$ from $L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$ and the second is to distinguish $L^\star$ from $L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$.

To address the first identifiability issue, we appeal to the previous literature on sparse-plus-low rank decompositions, which states that the matrices $S^\star$ and $L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$ are identifiable from their sum if the row and columns of the matrix $S^\star$ are sufficiently sparse and the matrix $L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$ is sufficiently low-rank with most of its entries non-zero and similar in magnitude (Candès et al., 2011; Chandrasekaran et al., 2011; Recht et al., 2010). Sparsity of $S^\star$ corresponds to small $d^\star$ in (13) so that no observed variable is directly connected to "many" other observed variables. Thus, we want $d^\star$ to be small so that no observed variable is directly connected to "many" other observed variables. Since the matrix $t\mathbf{1}_p\mathbf{1}_p^\top$ has equal entries and is rank one, the structural constraint on $L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$ can be interpreted as the number of latent variables being small (as compared to the ambient dimension p) with their effects spread across all the observed variables. To measure the "diffuseness" of the latent effects, we consider the following quantity for any linear subspace $Z \subseteq \mathbb{R}^p$ (Candès et al., 2011; Candès and Recht, 2012; Chandrasekaran et al., 2011, 2012): $\mu[Z] := \max_i \|\mathcal{P}_Z(\mathbf{e}_i)\|_2$, where $\mathcal{P}_Z$ is the projection onto the subspace Z and $\mathbf{e}_i$ is a standard coordinate basis. The quantity $\mu[Z]$ is also known as the "incoherence parameter" (Candès and Recht, 2012; Chandrasekaran et al., 2011). It measures how aligned the subspace Z is with respect to standard basis elements and is lower-bounded by $\sqrt{\dim(Z)/p}$ and upper-bounded by one. In our

setting, the relevant subspace is the row or column space of $L^\star$ and so we define

$$\mu^\star := \mu[\text{col-space}(L^\star)]. \tag{14}$$

Thus, a lower bound for $\mu^\star$ is $\sqrt{h/p}$, which is achieved when the effect of latent variables on the observed variables is equally spread out. A small value of $\mu^\star$ ensures the matrix $L^\star$ has a small rank and is far from being sparse.

To address the second identifiability issue of disentangling $L^\star$ and $t\mathbf{1}_p\mathbf{1}_p^\top$ from their sum, as described earlier, we want the deviation of the subspaces $T^\star$ and $\text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ to not be too large (i.e. the lower-bound condition in Assumption 3). This deviation can be conveniently measured by $\kappa^\star : -\|\mathcal{P}_{T^{\star\perp}}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2$ which is equivalent to $\|\mathcal{P}_{\text{col-space}(L^\star)^\perp}(1/\sqrt{p}\mathbf{1}_p)\|_F^2$.

Having these identifiability concerns in mind, we give stylized extremal graphical models and numerically check that the Hessian conditions in Assumptions 1–3 are satisfied for appropriate choice of parameters. Specifically, we set $p = 30$, $h = 1$ and specify the subgraph $\mathcal{G}_O = (E_O, O)$ among the observed variables to be an Erdős–Rënyi graph with edge probability $\tau \in \{0.001, 0.005\}$ and set $\Theta^\star_{ij}$ to 0.2 for every $(i, j) \in E_O$ and zero otherwise. We connect the latent variable to each observed variable and select the corresponding entries $\Theta^\star_{p+1,k}$ uniformly at random from the interval $[1/\sqrt{k}, 1.1/\sqrt{k}]$ for all $k \in O$. Notice that larger values of τ lead to larger sparsity parameter $d^\star$. We let $\omega = 0.003$ so that the largest angle between tangent spaces T' and $T^\star$ is less than 0.0005 degrees. Employing a numerical procedure described in Appendix D.1, we obtain a range of values of γ, α, ν that satisfy Assumptions 1–3. The values of α and ν that are computed using this procedure serve as a lower bound for the optimal α, ν, respectively. Indeed, an exciting direction for future research is to develop numerical or analytical techniques to precisely characterize the optimal values of α, ν. Table 1 illustrates $d^\star$ and the corresponding values of γ, α, ν that satisfy Assumptions 1–3. Examining Table 1, we can make two observations. First, for each value of τ, a larger range of γ results in smaller α and ν. Second, larger graph density (i.e., larger τ) reduces the range of values of γ that satisfy Assumption 1-3. These two observations are consistent with theory; see Remark 8.

τ	$d^\star$	γ	$\alpha \geq$	$\nu \leq$
0.001	1	(1.7,3.6)	0.91	0.004
0.001	1	(2.15,3)	1.14	0.150
0.005	2	(2.8,3.45)	1.26	0.007
0.005	2	(3,3.3)	1.3	0.04

Table 1: Different values of the edge probability τ, the maximum node degree (13), and the corresponding ranges of the regularization γ in (9) and values of α, ν that satisfy Assumptions 1–3.

4.4 Theorem statement

We now describe the performance of `eglatent` under suitable conditions on the quantities from the previous section. We state the theorem based on essential aspects of the conditions required for the success of our convex relaxation (i.e., the Hessian conditions) and omit complicated constants. We specify these constants in Appendix H. Our results depend on a second-order parameter $\xi > 0$ that determines the rate of convergence of a random vector X in the domain of attraction of a Hüsler–Reiss distribution to its limit, with larger values corresponding to faster convergence; see Appendix G.

Theorem 9 *Suppose that we have n independent and identically distributed samples in the domain of attraction of a latent Hüsler–Reiss model as described in Section 3.3 with second-order parameter $\xi > 0$ in Assumption 9 in Appendix G. Assume that there exists $\alpha > 0$, $\nu \in (0, 1]$, $\omega \in (0, 1)$ and the choice of the parameter γ so that the Hessian $\mathbb{I}^\star$ corresponding to this latent Hüsler–Reiss model satisfies Assumptions 1–3. Let $m := \max\{1, 1/\gamma\}$ and $\bar{m} := \max\{1, \gamma\}$. Let the effective sample size k be chosen such that $k < n^{2\xi/(2\xi+1)}$. Let $h := \operatorname{rank}(L^\star)$ be the true number of latent variables and $d^\star$ the maximum degree of the true graph structure among the observed variables conditioned on the latent variables as in* (13)*. Suppose:*

1. $k \gtrsim \frac{m^5 h d^{\star 2}}{\alpha^6} p^2 \log(p)$, *i.e., effective sample size is sufficiently large;*
2. $\lambda_n \sim \frac{m}{\nu}\sqrt{\frac{p^2\log(p)}{k}}$, *i.e., λ_n is appropriately chosen;*
3. $\sigma_{\min}(L^\star) \gtrsim \frac{m^4 \bar{m} h}{\nu\alpha^4}\sqrt{\frac{p^2\log(p)}{k}}$, *i.e., the minimum nonzero singular value of $L^\star$ is sufficiently bounded away from zero;*
4. $|S^\star_{ij}| \gtrsim \frac{m^3 \bar{m}\sqrt{h}}{\nu\alpha^2}\sqrt{\frac{p^2\log(p)}{k}}$ *for every (i, j) with $|S^\star_{ij}| > 0$, i.e., the minimum nonzero entry of $S^\star$ is sufficiently bounded away from zero.*

Then, the estimate $(\hat{S}, \hat{L})$ defined as the unique minimizer of `eglatent` *in* (9) *with empirical variogram in* (7) *satisfies*

$$\mathbb{P}\left(\operatorname{sign}(\hat{S}) = \operatorname{sign}(S^\star), \operatorname{rank}(\hat{L}) = \operatorname{rank}(L^\star), \|(\hat{S} - \hat{L}) - \tilde{\Theta}^\star\|_2 \lesssim \frac{m^3\sqrt{h}}{\nu\alpha^2}\sqrt{\frac{p^2\log(p)}{k}}\right) \geq 1 - \frac{1}{p}.$$

Remark 10 *The class of distributions X in the domain of attraction of a Hüsler–Reiss distribution is very large. For instance, the max-stable Hüsler–Reiss distribution is one member of this class. Note that since we are considering threshold exceedances, the right-hand side of* (1) *changes with the threshold u even if X is a max-stable distribution. Indeed, Engelke et al. (2022c, Proposition S.6) showed that the rate of convergence is governed by a second-order parameter ξ that can be chosen as any value in $(0, 1)$. In this case, the effective sample size k in Theorem 9 must satisfy $k = o(n^{0.66})$. Thus, in our simulations with max-stable distribution, we use $k = n^{0.65}$.*

We prove Theorem 9 in Appendix H. Due to the zero row-sum constraint in the `eglatent` estimator (9), the proof of Theorem 9 is more involved than the consistency analysis in Chandrasekaran et al. (2012). Specifically, we need additional technical arguments to deal with the dual parameter $t\mathbf{1}_p\mathbf{1}_p^\top$ that arises from the zero row-sum constraint. We highlight these technical arguments in Appendix D.1.

Theorem 9 essentially states that if Assumptions 1–3 hold, (λ, γ) are chosen appropriately, the effective sample size k is sufficiently large, the minimum nonzero singular value of the low-rank term $L^\star$ and the minimum nonzero entry of the sparse piece $S^\star$ are bounded away from zero, then, with high probability, `eglatent` provides accurate estimates for the subgraph among the observed variables, the number of latent variables, and a marginal extremal model.

The quantities (α, ν, ω) as well as the choices of the parameters λ_n and γ play a prominent role in the result. Indeed, larger values of α, ω, ν lead to a better conditioned Hessian $\mathbb{I}^\star$ around the tangent spaces $\Omega^\star$ and $T^\star \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)$. The better conditioning of the Hessian $\mathbb{I}^\star$ then results in less stringent requirements on sample complexity, the minimum nonzero singular value of $L^\star$, and the magnitude of the minimum nonzero entry of $S^\star$. Notice that the complexity of the true graph structure among the observed variables $d^\star$ and the true number of latent variables h appears explicitly in the bounds in Theorem 9. We also note the dependence on $d^\star$ and h is implicit in the dependence on α, ν and γ. Indeed, as larger $d^\star$ and h increases the dimension of the tangent space $\Omega^\star$ and $T^\star$, respectively, they result in smaller α, ν. Furthermore, as described in Remark 8 , the range of values of γ decrease with larger graph complexity and number of latent variables.

Remark 11 *Engelke et al. (2022c) prove that $k \geq \mathcal{O}(\log(p))$ suffices for consistent estimation of extremal graphical models without latent variables. Further, Chandrasekaran et al. (2012) prove that $k \geq \mathcal{O}(p)$ suffices for consistent estimation of Gaussian latent variable graphical model. According to Theorem 9, we require $k \geq \mathcal{O}(p^2 \log(p))$ in our setting. This requirement is determined by the deviation $\|\hat{\Gamma}_O - \Gamma^\star_O\|_2$, namely how fast the empirical variogram matrix $\hat{\Gamma}_O$ converges in* spectral norm *to the true variogram matrix $\Gamma^\star_O$. Engelke et al. (2022c) carried out extensive mathematical arguments to obtain the following concentration of the empirical variogram matrix in ℓ_∞ norm $\|\hat{\Gamma}_O - \Gamma^\star_O\|_\infty \leq \mathcal{O}(\sqrt{\log(p)/k})$. In our analysis, we use this result and the equivalence of norms relation*

$$\|\hat{\Gamma}_O - \Gamma^\star_O\|_2 \leq p\|\hat{\Gamma}_O - \Gamma^\star_O\|_\infty \leq \mathcal{O}(\sqrt{p^2 \log(p)/k})$$

to obtain a convergence rate in the spectral norm. We suspect that a tighter convergence result of $\|\hat{\Gamma}_O - \Gamma^\star_O\|_2 \leq \mathcal{O}(\sqrt{p/k})$ holds. Such a tighter convergence guarantee would then imply $k \geq \mathcal{O}(p)$ is sufficient for consistency guarantees of our estimator.

Finally, we should expect a more stringent sample size requirement for the latent extremal model than for the extremal model without latent variables. In particular, a larger sample size allows us to guarantee spectral norm consistency of the low-rank component, ensuring accurate estimates for the number of latent variables and their effects; see also the discussion in Chandrasekaran et al. (2012).

5. Experimental demonstrations

cnIn our numerical experiments, we use `eglatent` as a model selection procedure and perform a second refitting step on the selected model structure to estimate the model parameters; see Appendix I for details. Code to reproduce our results can be found at `https://github.com/sebastian-engelke/extremal_latent_learning`.

5.1 Synthetic simulations

We illustrate the utility of our method for recovering the subgraph among the observed variables and the number of latent variables on synthetic data. We compare the performance of our `eglatent` method to `eglearn` by Engelke et al. (2022c) for learning extremal graphical models. (In Appendix J.3, we provide comparisons with the Gaussian latent variable estimator in Chandrasekaran et al. (2012). As expected, our estimator is better at capturing dependency structure in the extremes and outperforms the Gaussian estimator.) To evaluate the accuracy of the estimated graphs with edges $\hat{E}$ relative to the true subgraph among the observed variables with edges $E = E_O$, we use the F-score

$$F = \frac{|E \cap \hat{E}|}{|E \cap \hat{E}| + \frac{1}{2}(|E^c \cap \hat{E}| + |E \cap \hat{E}^c|)}.$$

Larger F-scores thus indicate more accurate graph recovery.

5.1.1 Structure recovery

In order to evaluate the performance of our new method, we generate data from a random vector $X = (X_O, X_H)$ in the domain of attraction of a latent Hüsler–Reiss multivariate Pareto distribution Y with the precision matrix $\Theta^\star \in \mathbb{R}^{p+h \times p+h}$, p observed variables $O = \{1, 2, \ldots, p\}$ and h latent variables $H = \{p+1, \ldots, p+h\}$. We choose to simulate X from the Hüsler–Reiss max-stable distribution with the same precision matrix $\Theta^\star$, which is well-known to be in the domain of attraction of Y; see Resnick (2008) for details. The simulation can be done efficiently with the method in Dombry et al. (2016).

We specify the sub-graph $\mathcal{G}_O = (E_O, O)$ among the observed variables to be a cycle graph and set $\Theta^\star_{ij}$ to -2 for every $(i, j) \in E_O$ and zero otherwise. The latent variables are not connected in the joint graph, so $\Theta^\star_{ij} = 0$ for every $i, j \in H, i \neq j$. We connect each latent variable node $i \in H$ to every $k \in O$ satisfying $k = i - (p+1) + \zeta h$ for some positive integer ζ (thus every latent variable is connected to a distinct set of observed variables in the graph). The corresponding entries $\Theta^\star_{ik}$ in the precision matrix are chosen uniformly at random from the interval $[50/\sqrt{p+h}, 75/\sqrt{p+h}]$. Finally, we set the diagonal entries of $\Theta^\star$ to have the all-ones vector in its null space. Appendix J.1 shows results for a setting where the subgraph among the observed variables is generated according to an Erdős–Rényi graph.

We let $p = 30$, $h \in \{1, 2, 3\}$, and we set the number of marginal exceedances to $k = \lfloor n^{0.65} \rfloor$. Following Remark 10, this choice satisfies the assumptions of Theorem 9 since we simulate from a max-stable distribution. Altering k in a reasonable range does not change the

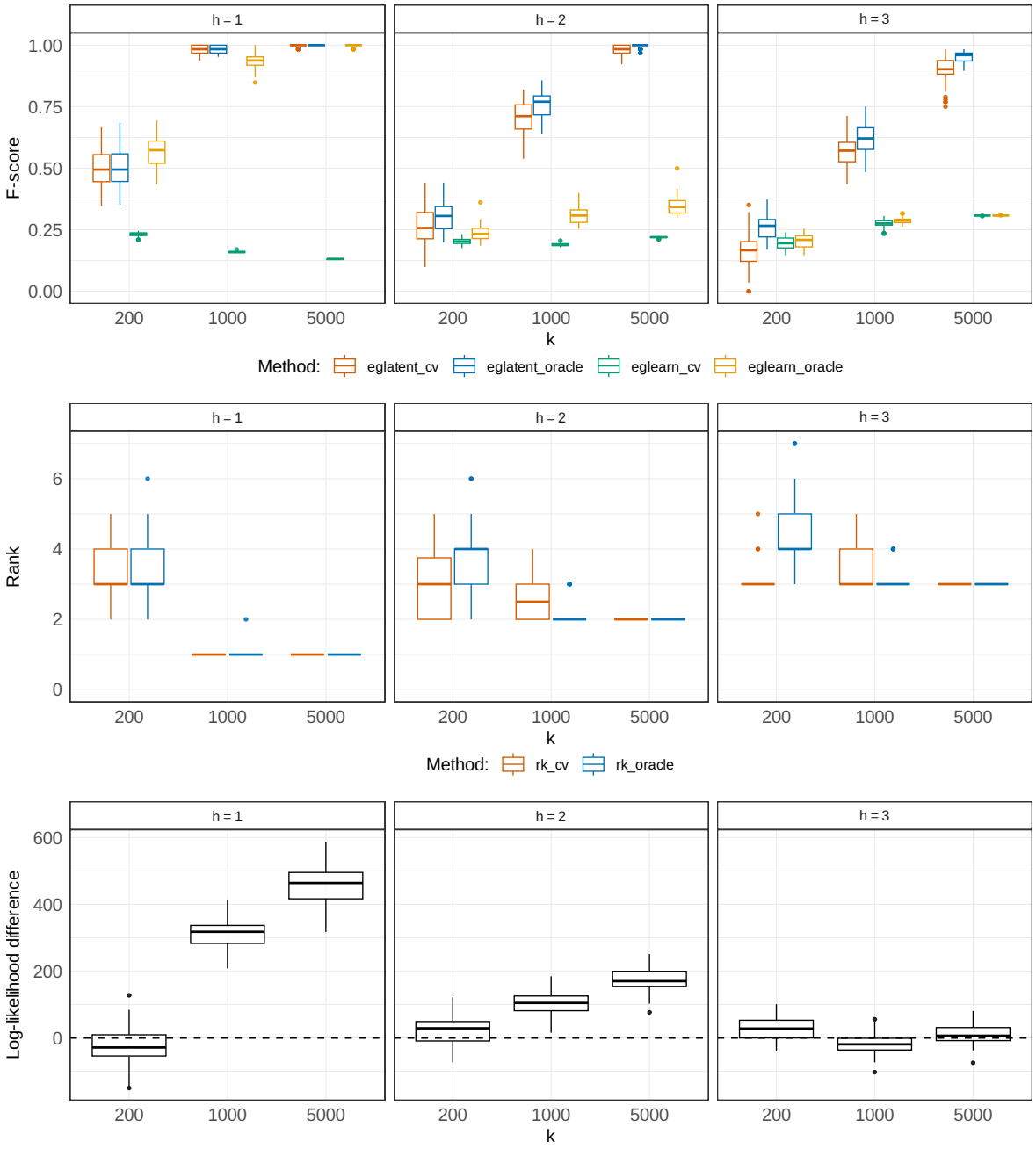

Figure 3: F-score (top row) and estimated number of latent variables (middle row) of `eglatent` method with the selection of the tuning parameter based on the oracle and validation on the F-score for the cycle graph with $h = 1, 2, 3$ latent variables and different effective sample sizes $k = 200, 1000, 5000$. The bottom row shows the difference between best `eglatent` and best `eglearn` log-likelihoods on the validation set.

qualitative results of the simulation study. A more detailed discussion of the choice of k can be found in the real data application in Section 5.2. We generate n samples from the max-stable Hüsler–Reiss distribution parameterized by $\Theta^\star$ so that we obtain $k \in \{200, 1000, 5000\}$ effective extreme samples. When deploying our `eglatent` estimator in (9), we fix $\gamma = 4$ to a reasonable default value; In Appendix J.2, we demonstrate the robustness of our results to different values of γ. Concerning the regularization parameter λ_n, which also appears in the `eglearn` method, in both methods, it is chosen either by validation likelihood on a separate dataset of size n or by an oracle approach maximizing the F-score for the sub-graph among observed variables.

Figure 3 summarizes the performance of the methods on 50 independent trials for the different sample sizes and different numbers of latent variables. We observe that our proposed approach outperforms `eglearn` in several ways. Indeed, the top row shows that the graph learned by `eglearn` only poorly recovers the graphical structure among observed variables. This reveals a limitation of this method, namely that in the presence of latent variables, the marginal graph of observed variables is dense and sparsity cannot be well detected by methods that ignore this fact. Clearly, this problem becomes more pronounced with a larger number of latent variables. On the other hand, our new `eglatent` method exploits the latent structure for learning the sparse graph among the observed variables conditional on the latent variables. It recovers the graphical structure among the observed variables increasingly well with a growing sample size. In fact, the results for the tuning parameter λ_n chosen through validation likelihood are almost as good as those based on the oracle. The middle row of Figure 3 shows that `eglatent` is able to identify the correct number of latent variables, especially for larger sample sizes.

We can also compare the model in terms of their likelihood on the validation data. Again, our `eglatent` method generally attains a better validation likelihood and is thus more representative of the data. As an exception, we observe that if the effective sample size is small ($k = 100$), then `eglearn` performs better. The reason is that `eglatent` is a more flexible model with more parameters to learn, and it therefore benefits more from additional data.

5.1.2 Robustness to zero latent variables

We now evaluate the performance of `eglatent` when there are no latent variables present and compare its performance to `eglearn`. We first specify a graph structure using a Barabási–Albert model denoted by $\mathrm{BA}(d, m)$, which is a preferential attachment model with d notes and a degree parameter m (Albert and Barabási, 2001). We set $d = 20$ and $m = 2$. We then define a Hüsler–Reiss precision matrix $\Theta^\star \in \mathbb{R}^{d\times d}$ with entries sampled uniformly at random from the interval $[-5, -2]$. The diagonal entries of $\Theta^\star$ are chosen so that it has the all-ones vector in its null space. We generate n samples from the max-stable Hüsler–Reiss distribution parameterized by $\Theta^\star$ such that there are $k = \lfloor n^{0.65} \rfloor = 200$ effective marginal extreme samples. We also generate a separate dataset of size n for validation. For the method `eglatent`, for each value of the regularization parameter $\gamma = 1, 4, 8, 20$ the regularization parameter λ_n is chosen based on the validation set. The regularization parameter λ_n in `eglearn` is chosen similarly.

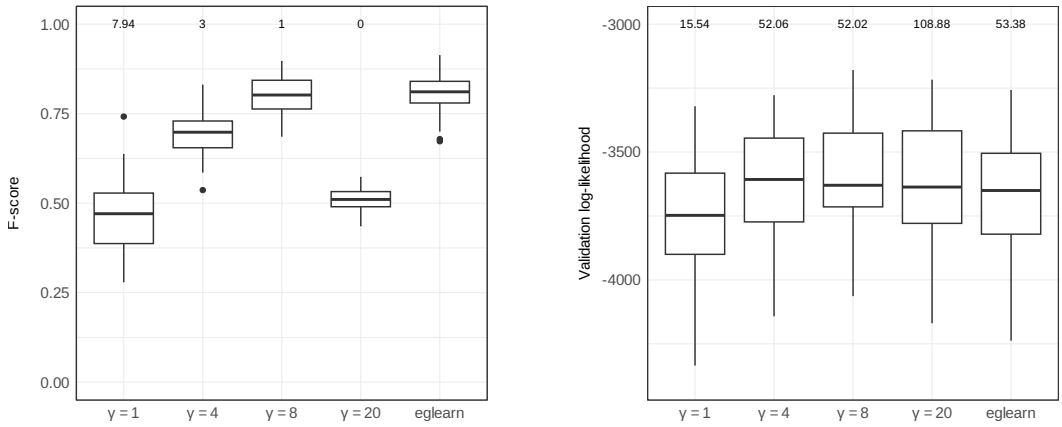

Figure 4: Left: F-score of `eglatent` for different regularization parameters $\gamma \in \{1, 4, 8, 20\}$ and `eglearn`; top axis shows the average number of estimated latent variables in `eglatent`. Right: the log-likelihood of the same methods evaluated on a validation data set; the top axis shows the average number of estimated edges in each model.

Figure 4 presents the F-scores and validation log-likelihood scores of `eglatent` as γ varies and for 50 independent trials. We also display the average numbers of edges and latent variables, as well as the performance of `eglearn`. As expected, larger values of γ lead to smaller estimates for the number of latent variables. We observe that when $\gamma = 4$, `eglatent` obtains an accurate graphical structure (F-score close to one) with a similar validation likelihood as `eglearn`. Here, `eglearn` yields a sparse graph since, unlike the previous settings, there are no unobserved confounding. Interestingly, the average number of estimated latent variables in this case is not close to zero. In particular, we observe that when γ is chosen so that `eglatent` yields nearly zero latent variables (i.e., $\hat{L} \approx 0$), the F-scores scores obtained by `eglatent` drop significantly. For such γ, our estimator (9) resembles the analog of the graphical lasso which is known to yield inaccurate models (Engelke et al., 2022c); see also Remark 6.

In summary, when the sample size is sufficiently large, `eglatent` yields a similar model fit and graph recovery as `eglearn` even when there are no latent variables. It is worth emphasizing that `eglatent` achieves this favorable performance by estimating some latent variables. This shows the robustness of our method to model misspecification.

5.2 Real data application

We apply our latent Hüsler–Reiss model to analyze large flight delays. We use a data set from the R package `graphicalExtremes` (Engelke et al., 2022a) with $p = 29$ airports in the southern U.S. shown in the left panel of Figure 5. Large flight delays cause huge financial losses and lead to congestion of critical airport infrastructure. Our method provides an improved model for the dependence of such excessive delays at different airports, and can eventually be used for stress testing of the system; see Hentschel et al. (2022) for details on this application. Unless otherwise noted, we fit the models in the whole dataset consisting of

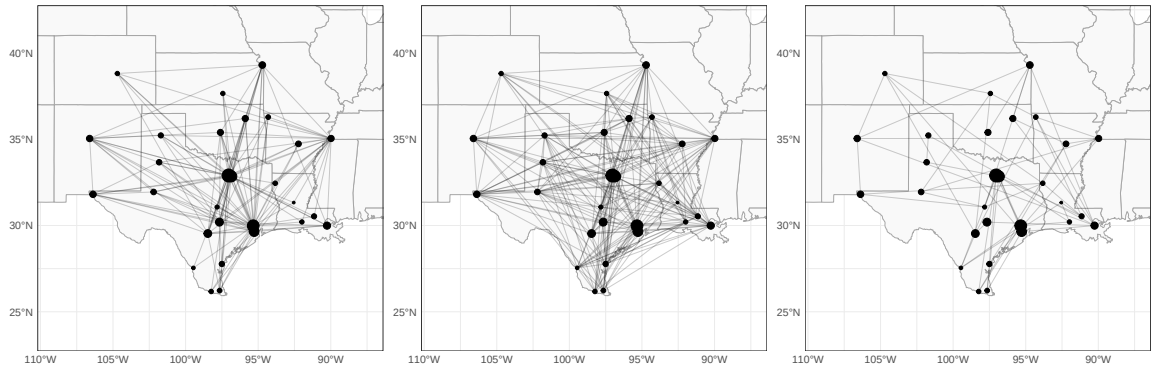

Figure 5: Airports in the Southern U.S. (dots) and flight connections, where the thickness of the nodes indicates the average number of daily flights at the airports. Left: flight connection graph with an edge between any pair of airports with daily flights. Center: estimated graph of optimal `eglearn` model. Right: estimated sub-graph corresponding to observed variables of optimal `eglatent` model.

$n = 3603$ observations from 2005-01-01 to 2020-12-31. We compare our `eglatent` method for latent Hüsler–Reiss models with the `eglearn` algorithm by (Engelke and Volgushev, 2022) that estimates a graphical structure without latent variables. We report here the results for the exceedance threshold of be $q = 0.90$ (i.e., $1 - k/n = 0.90$) resulting in $k = 360$ marginal exceedances for the computation of the empirical variogram $\hat{\Gamma}_O$; see Section 3.3.1. The latter is the input for the different structure learning methods. Different choices of the threshold, or equivalently, of k, are discussed below.

The left-hand side of Figure 6 shows the number of edges of `eglatent` and of `eglearn` as a function of the tuning parameter λ_n, where the parameter γ related to the latent variable selection in `eglatent` is fixed to the default choice $\gamma = 4$; different values of γ give similar results and are omitted here. We see that for both methods, larger values of λ_n result in sparser graphs. It is important to note that for `eglearn`, we count the edges of the usual estimated graph. For our `eglatent` method we count the edges of the residual graph among the observed variables. The latent graphs generally have fewer edges and are therefore more easily interpretable.

To compare the different model fits and to select the optimal value for the tuning parameter λ_n, we must compute the likelihood of the fitted models on an independent validation set. To this end, we split the data chronologically into five equally large folds and perform cross-validation by leaving one fold out (validation data) and fitting on the remaining four folds (training data). The results for model performance on the validation sets are then averaged. The right-hand side of Figure 6 shows the averaged log-likelihood values on the validation sets that were not used for model fitting. For both methods, we see that for too small values of λ_n, the graphs are too dense and overfit to the training data. In fact, for $\lambda_n = 0$, both models correspond to the fully connected graph whose performance (horizontal line) is much worse than the models enforcing sparsity. For too large values of λ_n, the graph becomes too sparse and the model is not flexible enough. Clearly, the latent model outperforms `eglearn`, indicating that latent variables are present in this data set. In this particular application,

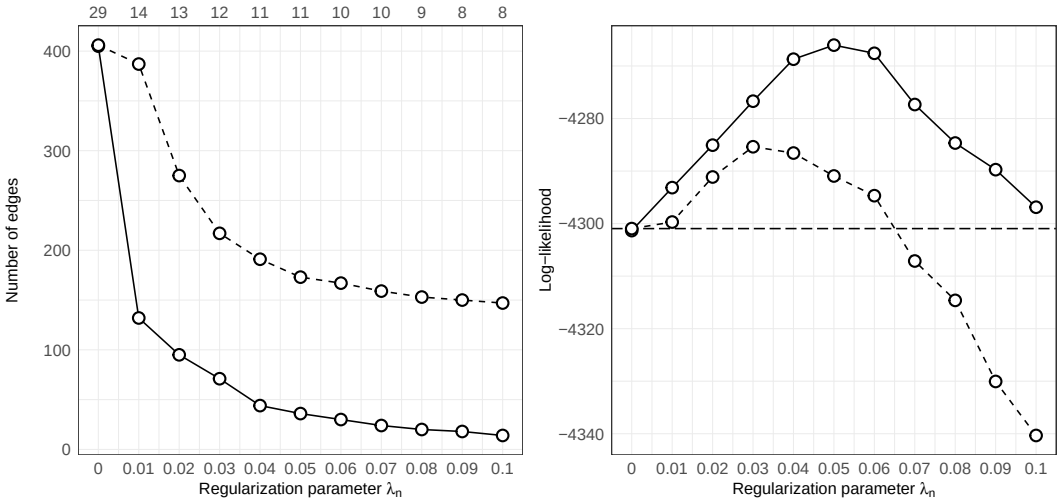

Figure 6: Left: number of edges of the estimated graph of `eglearn` (dashed line) and the estimated sub-graph of observed variables of `eglatent` (solid line) as functions of the regularization parameter ρ; top axis shows the number of latent variables in `eglatent`. Right: corresponding log-likelihoods; horizontal line is the validation log-likelihood of the fully connected graph.

they can be thought of as confounding factors such as meteorological variables or strikes in the aviation industry that affect many airports simultaneously.

Figure 5 compares the estimated graphs of `eglatent` (center) and `eglearn` (right) fitted on the whole data set, where the regularization parameter λ_n in both methods is chosen as the maximizers of the respective validation likelihoods. We observe that the latent graph is much sparser and therefore highlights more clearly certain features of the system. For instance, it seems that hubs, such as the Fort Worth International Airport in Dallas (the thickest point on the map), are more central in the graph since they have more connections than smaller airports.

The number of exceedances k used in the analysis, or equivalently, the probability threshold $q = 1 - k/n$, is a tuning parameter appearing in virtually all extreme value analyses. In theory and simulation studies with knowledge of the underlying distributions, there is an optimal choice of the asymptotic order of k compared to the sample size n; see for instance Remark 10. In real data, we typically neither know the data generating distribution nor the second-order parameter ξ that determines the rate of k in Theorem 9. Therefore, it is common practice to run the analysis for different choices of reasonable values of k and compare the results in terms of stability. In addition to the threshold $q = 0.90$ ($k = 360$), we rerun the above application with thresholds $q = 0.85$ ($k = 540$) and $q = 0.95$ ($k = 180$); see Appendix J.4 for the results. Similarly to Figure 6, Figures 9 and 10 show that also for these threshold choices, `eglatent` outperforms `eglearn` significantly. Moreover, Figure 11 compares the different estimated graphs among the observed variables for the three thresholds. We see that the results are very stable and the graphs only have a few edges that differ.

6. Future work

Our work on latent variables in the analysis of extremal dependence opens several future research directions. First, as described in Section 4.4, our sample size requirement is driven by a spectral norm concentration result on the empirical variogram matrix. This result was derived by translating the ℓ_∞ concentration result of Engelke et al. (2022c) to the spectral norm setting using equivalence of norms. To obtain tighter convergence results, one must obtain direct concentration bounds on the spectral norm; such a result would be of independent interest in the multivariate extremes literature. Second, solving `eglatent` can be challenging for large problems. Building on the work of Ma et al. (2012) in the Gaussian setting, faster solvers can be developed using alternating direction method of multipliers (Boyd et al., 2011). Moreover, we observed in Section 5.1.2 that `eglatent` estimates a few latent variables to accurately recover the underlying graphical structure when there are no latent variables present. It would be of interest to develop a theoretical justification for this phenomenon. Also additional structure on the dependency structures among the observed and latent variables, such as multivariate total positivity of order 2 (Röttger et al., 2023b; Rodríguez and Röttger, 2024) or colored graphs (Röttger et al., 2023a), may be exploited to develop more powerful extremal graphical models with latent variables. The recent connection between extremal graphical models and graphical models for Lévy processes could allow us to use `eglatent` also for modeling latent variables in stochastic processes dependence structure (Engelke et al., 2024b).

Acknowledgments

We thank the referees for their valuable comments that improved the paper. We thank Nicola Gnecco and Manuel Hentschel for their help with creating the figures in this paper. The authors acknowledge funding from the Swiss National Science Foundation (Sebastian Engelke), NSF grant DMS-2413074 and the Royalty Research Fund at the University of Washington (Armeen Taeb).

Appendix A. Useful lemmas for proving Theorem 5

Our analysis of Theorem 5 relies on some lemmas.

Lemma 12 *Let $A \in \mathbb{S}^d$ and $B \in \mathbb{S}^d$ be two symmetric matrices with $A+B$ being nonsingular and row/column spaces of A and B being orthogonal to one another. Then, $(A+B)^{-1} = A^+ + B^+$.*

Proof [Proof of Lemma 12] Let $U_A D_A U_A^T$ and $U_B D_B U_B^T$ be the reduced SVD of A and B. Then, since $A+B$ is non-singular, and the subspaces spanned by the columns of U_A and U_B are orthogonal, we have that (U_A, U_B) forms an orthogonal matrix. Therefore,

$$(A+B) = \begin{pmatrix} U_A & U_B \end{pmatrix} \begin{pmatrix} D_A & 0 \\ 0 & D_B \end{pmatrix} \begin{pmatrix} U_A & U_B \end{pmatrix}^T,$$

and thus $(A+B)^{-1} = U_A D_A^{-1} U_A^T + U_B D_B^{-1} U_B^T = A^+ + B^+$. ■

Lemma 13 *Suppose that $UU^T MUU^T = M$. Then, $U(U^T MU)^{-1}U^T = M^+$.*

Proof [Proof of Lemma 13] Let UDU^T be the reduced-SVD of M. Then, $U(U^M U)^{-1}U^T = UD^{-1}U^T$, which is equivalent to M^+. ■

Lemma 14 *Let $\tilde{\Pi} = (I_p - \mathbf{1}_p\mathbf{1}_p^\top/p)$. Let $\Theta = \begin{pmatrix} \Theta_O & \Theta_{OH} \\ \Theta_{HO} & \Theta_H \end{pmatrix} \in \mathbb{R}^{d\times d}$ with $\Theta_O \in \mathbb{R}^{p\times p}$, $\Theta_H \in \mathbb{R}^{h\times h}$ and $d = h + p$. Suppose Θ is a positive semi-definite matrix with its null-space being the span of the all-ones vector. Then:*

$$\tilde{\Pi}(\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO})\tilde{\Pi} = \Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO}.$$

Proof Since $\Theta\mathbf{1}_d = 0$, we have

$$\Theta_O\mathbf{1}_p + \Theta_{OH}\mathbf{1}_h = 0, \tag{15}$$
$$\Theta_{HO}\mathbf{1}_p + \Theta_H\mathbf{1}_h = 0. \tag{16}$$

Consider $\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO}$, we have

$$\begin{aligned}(\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO})\mathbf{1}_p &= \Theta_O\mathbf{1}_p - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO}\mathbf{1}_p, \\ &\overset{by(16)}{=} \Theta_O\mathbf{1}_p + \Theta_{OH}{\Theta_H}^{-1}(\Theta_H\mathbf{1}_h), \\ &= \Theta_O\mathbf{1}_p + \Theta_{OH}\mathbf{1}_h \overset{by(15)}{=} 0.\end{aligned}$$

Thus, $\mathbf{1}_p \in \ker(\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO})$. To complete the proof, we will show that $\dim(\ker(\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO})) = 1$. Suppose there exist non-zero vector $v \in \ker(\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO})$, and let $u = {\Theta_H}^{-1}\Theta_{HO}v$. Since $v \neq 0$, $u \neq 0$, and then it follows that $\Theta_O v - \Theta_{OH}u = 0$ and $\Theta_{HO}v - \Theta_H u = 0$ yielding $\Theta^\star \begin{pmatrix} v \\ -u \end{pmatrix} = 0$. Since $\begin{pmatrix} v \\ -u \end{pmatrix} \in \ker(\Theta)$, $v = \alpha'\mathbf{1}_p$ for some $\alpha' \in R$, which implies that $\dim(\ker(\Theta_O - \Theta_{OH}{\Theta_H}^{-1}\Theta_{HO})) = 1$. ■

Lemma 15 *Let $\tilde{\Pi} = I_p - \mathbf{1}_p\mathbf{1}_p^\top/p$ and $\Pi = I_d - \mathbf{1}_d\mathbf{1}_d^T/d$ with $d = p + h$. For any matrix $M \in \mathbb{R}^{d\times d}$, $\tilde{\Pi}(\Pi M\Pi)_{1:p,1:p}\tilde{\Pi} = \tilde{\Pi}M_{1:p,1:p}\tilde{\Pi}$*

Proof Note that $(\Pi M\Pi)_{1:p,1:p} = \begin{pmatrix} I_p & \mathbf{0} \end{pmatrix} \Pi M\Pi \begin{pmatrix} I_p \\ \mathbf{0} \end{pmatrix}$. Then it follows that

$$\tilde{\Pi}(\Pi M\Pi)_{1:p,1:p}\tilde{\Pi} = \tilde{\Pi}\begin{pmatrix} I_p & \mathbf{0} \end{pmatrix} \Pi M\Pi \begin{pmatrix} I_p \\ \mathbf{0} \end{pmatrix} \tilde{\Pi} = \begin{pmatrix} \tilde{\Pi} & \mathbf{0} \end{pmatrix} \Pi M\Pi \begin{pmatrix} \tilde{\Pi} \\ \mathbf{0} \end{pmatrix}. \tag{17}$$

Notice that:

$$\begin{aligned}\begin{pmatrix}\tilde{\Pi} & \mathbf{0}\end{pmatrix}\Pi &= \begin{pmatrix}(I_p - \mathbf{1}_p\mathbf{1}_p^\top/p)(I_p - \mathbf{1}_p\mathbf{1}_p^\top/d) & (I_p - \mathbf{1}_p\mathbf{1}_p^\top/p)\mathbf{1}_p/d\end{pmatrix} \\ &= \begin{pmatrix}(I_p - \mathbf{1}_p\mathbf{1}_p^\top/p)(I_p - \mathbf{1}_p\mathbf{1}_p^\top/p + \mathbf{1}_p\mathbf{1}_p^\top/(p) - \mathbf{1}_p\mathbf{1}_p^\top/d) & \mathbf{0}\end{pmatrix} \\ &= \begin{pmatrix}\tilde{\Pi}(\tilde{\Pi} + \mathbf{1}_p\mathbf{1}_p^\top/(p) - \mathbf{1}_p\mathbf{1}_p^\top/d) & \mathbf{0}\end{pmatrix} = \begin{pmatrix}\tilde{\Pi} & \mathbf{0}\end{pmatrix}.\end{aligned} \tag{18}$$

Putting (17) and (18) together, we have the desired result.

■

Appendix B. Proof of Theorem 5

Proof [Proof of Theorem 5] For notational simplicity, we let $M = -\Gamma^\star/2$. Let $\Pi = I_d - \mathbf{1}_d\mathbf{1}_d^T/d$. We have from Hentschel et al. (2022) that $(\Pi M \Pi)^+ = \Theta^\star$ or equivalently $\Pi M \Pi = (\Theta^\star)^+$. Since $\Theta^\star$ has zero row/column sums and thus its row/column spaces are orthogonal to the all-ones vector, we have by Lemma 12 that for any $t > 0$, $(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1} = \Theta^{\star +} + (t\mathbf{1}_d\mathbf{1}_d^T)^+ = \Theta^{\star +} + \frac{1}{td^2}(\mathbf{1}_d\mathbf{1}_d^T)$. As $\Pi\mathbf{1}_d\mathbf{1}_d^T\Pi = 0$, we have that:

$$\Pi M \Pi = \Pi(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}\Pi.$$

The equation above implies $\tilde{\Pi}[\Pi M \Pi]_{1:p,1:p}\tilde{\Pi} = \tilde{\Pi}[\Pi(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}\Pi]_{1:p,1:p}\tilde{\Pi}$. Using Lemma 15, we have that:

$$\tilde{\Pi} M_{1:p,1:p}\tilde{\Pi} = \tilde{\Pi}[(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}]_{1:p,1:p}\tilde{\Pi}. \tag{19}$$

We will now analyze the term $[(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}]_{1:p,1:p}$ inside (19). From Schur's complement, we have that:

$$\begin{aligned}&[(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}]_{1:p,1:p} = \\ &\left[\Theta^\star_O + t\mathbf{1}_p\mathbf{1}_p^T - (\Theta^\star_{OH} + t\mathbf{1}_p\mathbf{1}_h^T)(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}(\Theta^\star_{HO} + t\mathbf{1}_h\mathbf{1}_p^T)\right]^{-1}.\end{aligned} \tag{20}$$

By the Woodbury inversion lemma, we have that:

$$(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1} = (\Theta^\star_H)^{-1} - (\Theta^\star_H)^{-1}\mathbf{1}_h\left(\frac{1}{t} + \mathbf{1}_h^T(\Theta^\star_H)^{-1}\mathbf{1}_h\right)^{-1}\mathbf{1}_h^T(\Theta^\star_H)^{-1}. \tag{21}$$

Plugging the result of (21) into (20), we have that:

$$\begin{aligned}&[(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}]_{1:p,1:p} \\ &= \Theta^\star_O + t\mathbf{1}_p\mathbf{1}_p^T - (\Theta^\star_{OH} + t\mathbf{1}_p\mathbf{1}_h^T)(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}(\Theta^\star_{HO} + t\mathbf{1}_h\mathbf{1}_p^T) = A + B + C\end{aligned}$$

where

$$\begin{aligned}A &= \Theta^\star_O - \Theta^\star_{OH}(\Theta^\star_H)^{-1}\Theta^\star_{HO}, \\ B &= t\mathbf{1}_p\mathbf{1}_p^T + t^2\mathbf{1}_p\mathbf{1}_h^T(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}\mathbf{1}_h\mathbf{1}_p^T, \\ C &= t\mathbf{1}_p\mathbf{1}_h^T(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}\Theta^\star_{HO} + t\Theta^\star_{OH}(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}\mathbf{1}_h\mathbf{1}_p^T \\ &\quad + \Theta^\star_{OH}(\Theta^\star_H)^{-1}\mathbf{1}_h\left(\frac{1}{t} + \mathbf{1}_h^T(\Theta^\star_H)^{-1}\mathbf{1}_h\right)^{-1}\mathbf{1}_h^T(\Theta^\star_H)^{-1}\Theta^\star_{HO}.\end{aligned}$$

From Lemma 14, we have that: $\tilde{\Pi}A\tilde{\Pi} = A$. Furthermore, notice that B lies in the all-ones subspace, i.e. $\tilde{\Pi}B\tilde{\Pi} = 0$ and is a positive semi-definite matrix for $t > 0$. Thus, the matrix $A + B$ is invertible. Notice

$$\lim_{t\to 0} t\mathbf{1}_p\mathbf{1}_h^T(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}\Theta^\star_{HO} = \lim_{t\to 0} t\mathbf{1}_p\mathbf{1}_h^T(\Theta^\star_H)^{-1}\Theta^\star_{HO} = 0,$$
$$\lim_{t\to 0} t\Theta^\star_{OH}(\Theta^\star_H + t\mathbf{1}_h\mathbf{1}_h^T)^{-1}\mathbf{1}_h\mathbf{1}_p^T = \lim_{t\to 0} t\Theta^\star_{OH}(\Theta^\star_H)^{-1}\mathbf{1}_h\mathbf{1}_p^T = 0,$$
$$\lim_{t\to 0} \Theta^\star_{OH}(\Theta^\star_H)^{-1}\mathbf{1}_h\left(\frac{1}{t} + \mathbf{1}_h^T(\Theta^\star_H)^{-1}\mathbf{1}_h\right)^{-1}\mathbf{1}_h^T(\Theta^\star_H)^{-1}\Theta^\star_{HO} = \lim_{t\to 0} t\Theta^\star_{OH}(\Theta^\star_H)^{-1}\mathbf{1}_h(\Theta^\star_H)^{-1}\Theta^\star_{HO} = 0,$$

so that $\lim_{t\to 0} C = 0$. Notice on the other hand that $\lim_{t\to 0} A + B \neq 0$. By the Woodbury inversion lemma, we have that: $(A + B + C)^{-1} = (A + B)^{-1} - (A + B)^{-1}C(I + (A + B)^{-1}C)^{-1}(A + B)^{-1}$. Thus:

$$\lim_{t\to 0} \tilde{\Pi}(A + B + C)^{-1}\tilde{\Pi} = \tilde{\Pi}\lim_{t\to 0}(A + B)^{-1}\tilde{\Pi} - \lim_{t\to 0}\tilde{\Pi}(A + B)^{-1}C(I + A^{-1}C)^{-1}A^{-1}\tilde{\Pi}.$$

Since $\lim_{t\to 0} C = 0$, we have that:

$$\lim_{t\to\infty} \tilde{\Pi}[(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}]_{1:p,1:p}\tilde{\Pi} = \lim_{t\to 0}\tilde{\Pi}(A + B + C)^{-1}\tilde{\Pi} = \lim_{t\to 0}\tilde{\Pi}(A + B)^{-1}\tilde{\Pi} = \tilde{\Pi}A^+\tilde{\Pi} = A^+.$$

Here, the second equality follows from noting that the row/column spaces of A and B are orthogonal to one another and so by Lemma 12, $(A + B)^{-1} = A^+ + B^+$. Furthermore, since B is a multiple of all-ones matrix, $\tilde{\Gamma}B^+\tilde{\Gamma} = 0$. The last equality follows from Lemma 14. Noting that $M_{1:p,1:p} = -\Gamma^\star_O/2$ and plugging in A^+ for $\tilde{\Pi}[(\Theta^\star + t\mathbf{1}_d\mathbf{1}_d^T)^{-1}]_{1:p,1:p}\tilde{\Pi}$ in (19), we conclude that:

$$(\tilde{\Pi}(-\Gamma^\star_O/2)\tilde{\Pi})^+ = \Theta^\star_O - \Theta^\star_{OH}(\Theta^\star_H)^{-1}\Theta^\star_{HO}.$$

Taking pseudo-inverses of both sides, we have the desired result. In Lemma 14, we also showed that $\Theta^\star_O - \Theta^\star_{OH}(\Theta^\star_H)^{-1}\Theta^\star_{HO} = \tilde{\Pi}(\Theta^\star_O - \Theta^\star_{OH}(\Theta^\star_H)^{-1}\Theta^\star_{HO})\tilde{\Pi}$.

■

Appendix C. Arriving at estimator (9)

Recall that $\tilde{\Theta}^\star = (\tilde{\Pi}(-\Gamma^\star_O/2)\tilde{\Pi})^+$, where $\tilde{\Pi} = UU^T$. Furthermore, the null-space of $\tilde{\Theta}^\star$ is the subspace span($\mathbf{1}_p\mathbf{1}_p^\top$). In other words, $UU^T\tilde{\Theta}^\star UU^T = \tilde{\Theta}^\star$. We arrive at our estimator by noting that $\tilde{\Theta}^\star$ is the unique minimizer of the convex program:

$$\begin{aligned}\hat{\Theta} = \underset{\Theta\in\mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det\left(U^T\Theta U\right) - \frac{1}{2}\operatorname{tr}(\Theta\Gamma^\star_O), \\ \text{s.t} \quad & \Theta \succeq 0 \ , \ \ \Theta\mathbf{1}_p = 0.\end{aligned} \tag{22}$$

To see why that is, first note that the constraint $\Theta \succeq 0$ can be removed since the log-det function forces $U^T\Theta U$ to be positive definite and together with the constraint $\Theta\mathbf{1}_p$ forces

$\Theta \succeq 0$ and additionally $UU^T\Theta UU^T = \Theta$. Note that $\mathrm{tr}(\Theta\Gamma^\star_O) = \mathrm{tr}(UU^T\Theta UU^T\Gamma^\star_O) = \mathrm{tr}(\Theta UU^T\Gamma^\star_O UU^T)$. Thus, an equivalent optimization to (22) is

$$\begin{aligned} \hat{\Theta} = \underset{\Theta\in\mathbb{S}^p}{\text{argmin}} \quad & -\log\det\left(U^T\Theta U\right) - \frac{1}{2}\mathrm{tr}(\Theta UU^T\Gamma^\star_O UU^T), \\ \text{s.t} \quad & \Theta \in \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp. \end{aligned} \tag{23}$$

Using Lagrangian duality theory, we have that $\hat{\Theta}$ must satisfy for some $t \in \mathbb{R}$

$$-U(U^T\hat{\Theta}U^{-1})U^T - \frac{1}{2}UU^T\Gamma^\star_O UU^T + t\mathbf{1}_p\mathbf{1}_p^\top = 0.$$

Note that $t = 0$ since the first two terms live in the space spanned by the columns of U and the last term lies in the orthogonal subspace. Similarly, $-U(U^T\hat{\Theta}U^{-1})U^T - \frac{1}{2}UU^T\Gamma^\star_O UU^T = 0$. Since $UU^T\hat{\Theta}UU^T = \hat{\Theta}$, we appeal to Lemma 13 to conclude that $\hat{\Theta}^+ = -\frac{1}{2}\frac{1}{2}UU^T\Gamma^\star_O UU^T$. Some simple manipulations allow us to conclude that $\hat{\Theta} = \tilde{\Theta}^\star$.

Appendix D. Useful lemmas for proof of consistency

Our analysis will depend on the following quantities for any pair of subspaces $\Omega, T \subseteq \mathbb{R}^{p\times p}$:

$$\theta(\Omega) := \max_{N\in\Omega, \|N\|_\infty=1} \|N\|_2 \quad ; \quad \xi(T) := \max_{N\in T, \|N\|_2=1} \|N\|_\infty.$$

When $\Omega = \Omega^\star$ and $T = T^\star$, these quantities are closely connected to the maximal degree $d^\star$ and the incoherence parameter $\mu^\star$ (defined in Section 4.1). In particular, Chandrasekaran et al. (2012) showed that $\mu(\Omega^\star) \in [0, d^\star]$ and $\xi(T^\star) \in [\mu^\star, 2\mu^\star]$.

D.1 Some auxillary lemmas

Lemma 16 (Lemma 3.1 of Chandrasekaran et al. (2012)) *For any tangent spaces T_1, T_2 of same dimension with $\rho(T_1, T_2) < 1$, we have that: $\xi(T_2) \leq \frac{\xi(T_1)+\rho(T_1,T_2)}{1-\rho(T_1,T_2)}$.*

Lemma 17 *Consider a tangent space T' of a symmetric matrix with $\rho(T^\star, T') \leq \omega$ with $\omega < 1$. Let $\mathcal{C}'$ and $\mathcal{C}^\star$ be the column spaces that form the tangent spaces T' and $T^\star$ respectively. Then, we have that: $\|\mathcal{P}_{\mathcal{C}'} - \mathcal{P}_{\mathcal{C}^\star}\|_2 \leq \omega$.*

Proof [Proof of Lemma 17] Since $\omega < 1$, $T^\star$ and T' are of the same dimension. Let $\sigma_s(\cdot)$ be the s-th largest singular value of the input matrix. Notice that

$$\begin{aligned} \|\mathcal{P}_{\mathcal{C}'} - \mathcal{P}_{\mathcal{C}^\star}\|_2 = \|\mathcal{P}_{\mathcal{C}'^\perp} - \mathcal{P}_{\mathcal{C}^{\star\perp}}\|_2 = \sqrt{1 - \sigma_{p-k}\left(\mathcal{P}_{\mathcal{C}'^\perp}\mathcal{P}_{\mathcal{C}^{\star\perp}}\right)^2} &= \sqrt{1 - \sigma_{(p-k)^2}\left(\mathcal{P}_{T'^\perp}\mathcal{P}_{T^{\star\perp}}\right)} \\ &\leq \sqrt{1 - \sigma_{(p-k)^2}\left(\mathcal{P}_{T'^\perp}\mathcal{P}_{T^{\star\perp}}\right)^2} \\ &= \|\mathcal{P}_{T'^\perp} - \mathcal{P}_{T^{\star\perp}}\|_2 = \|\mathcal{P}_{T'} - \mathcal{P}_{T^\star}\|_2. \end{aligned}$$

■

D.2 Lemmas to account for the zero row-sum constraint

To deal with the additional dual parameter $t\mathbf{1}_p\mathbf{1}_p^\top$ introduced by the zero row-sum constraint $(S-L)\mathbf{1}_p$, our analysis requires the following lemmas.

Lemma 18 *Let $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathbb{R}^p$ be a pair of subspaces. Then, for any $z \in \mathbb{R}^p$:*

$$\max_{v\in\mathcal{C}_1\oplus\mathcal{C}_2,\|v\|_2=1}\langle z,v\rangle \leq 2\min\left\{\max_{v\in\mathcal{C}_1,\|v\|_2=1}\langle z,v\rangle, \max_{v\in\mathcal{C}_2,\|v\|_2=1}\langle z,v\rangle\right\} + \max\left\{\max_{v\in\mathcal{C}_1,\|v\|_2=1}\langle z,v\rangle, \max_{v\in\mathcal{C}_2,\|v\|_2=1}\langle z,v\rangle\right\}.$$

Proof [Proof of Lemma 18] Suppose without loss of generality that $\max_{u_1\in\mathcal{C}_1,\|u_1\|_2=1} u_1^T z \leq \max_{u_2\in\mathcal{C}_2,\|u_2\|_2=1} u_2^T z$. Thus

$$\begin{aligned}
\max_{v\in\mathcal{C}_1\oplus\mathcal{C}_2,\|v\|_2=1}\langle z,v\rangle &= \max_{\substack{u_1\in\mathcal{C}_1,u_2\in\mathcal{C}_2,\|u_1\|_2=\|u_2\|_2=1\\ v=c_1u_1+c_2u_2}} |v^Tz|/\|v\|_2,\\
&= \max_{\substack{u_1\in\mathcal{C}_1,u_2\in\mathcal{C}_2,\|u_1\|_2=\|u_2\|_2=1\\ u_3=u_2-(u_2^Tu_1)u_1\\ v=c_1u_1+c_2u_3}} |v^Tz|/\|v\|_2,\\
&\leq \max_{\substack{u_1\in\mathcal{C}_1,u_2\in\mathcal{C}_2,\|u_1\|_2=\|u_2\|_2=1\\ u_3=u_2-(u_2^Tu_1)u_1\\ v=c_1u_1+c_2u_3}} \frac{|c_1|}{\sqrt{c_1^2+c_2^2}}|u_1^Tz| + \frac{|c_2|}{\sqrt{c_1^2+c_2^2}}|u_3^Tz|,\\
&\leq \max_{u_1\in\mathcal{C}_1,\|u_1\|_2=1} 2|u_1^Tz| + \max_{u_2\in\mathcal{C}_2,\|u_2\|_2=1}|u_2^Tz|.
\end{aligned}$$

■

Lemma 19 *Let $Z \in T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\rho(T',T^\star)\leq\omega$ and $\|Z\|_2=1$. Then, $1+2(\kappa^\star+\omega)\geq\|\mathcal{P}_{T'}(Z)\|_2\geq 1-2(\kappa^\star+\omega)$ and thus $\|\mathcal{P}_{T'^\perp}(Z)\|_2\leq 2(\kappa^\star+\omega)$.*

Proof [Proof of Lemma 19] Note that $\|Z\|_2+\|\mathcal{P}_{T'^\perp}(Z)\|_2\geq\|\mathcal{P}_{T'}(Z)\|_2\geq\|Z\|_2-\|\mathcal{P}_{T'^\perp}(Z)\|_2$. Let T' be a tangent space with associated row and column spaces $\mathcal{C}'$ and $\mathcal{R}'$. Let $\tilde{C}=\mathcal{C}'\oplus\mathrm{span}(\mathbf{1}_p)$ and $\tilde{R}=\mathcal{R}'\oplus\mathrm{span}(\mathbf{1}_p)$. Since $Z\in T'\oplus\mathrm{span}(\mathbf{1}_p)$, it is straightforward to show that $Z=\mathcal{P}_{\tilde{\mathcal{C}}}Z\mathcal{P}_{\tilde{\mathcal{R}}^\perp}+Z\mathcal{P}_{\tilde{\mathcal{R}}}$. Therefore, we have that $\mathcal{P}_{T'^\perp}(Z)=\mathcal{P}_{\mathcal{C}'^\perp}\left[\mathcal{P}_{\tilde{\mathcal{C}}}Z\mathcal{P}_{\tilde{\mathcal{R}}^\perp}+Z\mathcal{P}_{\tilde{\mathcal{R}}}\right]\mathcal{P}_{\mathcal{R}'^\perp}$. Thus, $\|\mathcal{P}_{T'^\perp}(Z)\|_2\leq\|\mathcal{P}_{\mathcal{C}'^\perp}\mathcal{P}_{\tilde{\mathcal{C}}}\|_2+\|\mathcal{P}_{\tilde{\mathcal{R}}}\mathcal{P}_{\mathcal{R}'^\perp}\|_2$. Letting $\mathcal{C}_1=\mathcal{C}'$ and $\mathcal{C}_2=\mathrm{span}(\mathbf{1}_p)$, we appeal to Lemma 18 to conclude that:

$$\begin{aligned}
\max_{v\in\tilde{\mathcal{C}},\|v\|_2=1}\|\mathcal{P}_{\mathcal{C}'^\perp}(v)\|_2 &\leq \max_{\substack{z\in\mathcal{C}'^\perp\\ \|z\|_2=1}}\max_{u_1\in\mathcal{C}',\|u_1\|_2=1} 2|\langle z,u_1\rangle| + \max_{\substack{z\in\mathcal{C}'^\perp\\ \|z\|_2=1}}\max_{u_2\in\mathrm{span}(\mathbf{1}),\|u_2\|_2=1}|\langle z,u_2\rangle|\\
&= \|\mathcal{P}_{\mathcal{C}'^\perp}(\mathbf{1}_p/\sqrt{p})\|_2.
\end{aligned}$$

Again, appealing to Lemma 18,

$$\begin{aligned}\max_{v\in\mathcal{C}'^{\perp},\|v\|_2=1}\|\mathcal{P}_{\tilde{\mathcal{C}}}(z)\|_2 &\leq \max_{\substack{z\in\mathcal{C}'^{\perp}\\ \|z\|_2=1}}\max_{u_1\in\mathcal{C}',\|u_1\|_2=1} 2|\langle z,u_1\rangle| + \max_{\substack{z\in\mathcal{C}'^{\perp}\\ \|z\|_2=1}}\max_{u_2\in\mathrm{span}(\mathbf{1}),\|u_2\|_2=1}|\langle z,u_2\rangle| \\ &= \|\mathcal{P}_{\mathcal{C}'^{\perp}}(\mathbf{1}_p/\sqrt{p})\|_2.\end{aligned}$$

So we have concluded that $\|\mathcal{P}_{\mathcal{C}'^{\perp}}\mathcal{P}_{\tilde{\mathcal{C}}}\|_2 \leq \|\mathcal{P}_{\mathcal{C}'^{\perp}}(\mathbf{1}/\sqrt{p})\|_2$. Thus, appealing to Lemma 17 , $\|\mathcal{P}_{\mathcal{C}'^{\perp}}\mathcal{P}_{\tilde{\mathcal{C}}}\|_2 \leq \kappa^\star+\omega$. Similarly, we have that: $\|\mathcal{P}_{\mathcal{R}'^{\perp}}\mathcal{P}_{\tilde{\mathcal{R}}}\|_2 \leq \|\mathcal{P}_{\mathcal{R}'^{\perp}}(\mathbf{1}/\sqrt{p})\|_2$ and thus $\|\mathcal{P}_{\mathcal{R}'^{\perp}}\mathcal{P}_{\tilde{\mathcal{R}}}\|_2 \leq \kappa^\star+\omega$. Putting things together, we have the desired bound. ■

Lemma 20 *Let $T'\subseteq\mathbb{R}^{p\times p}$ be a tangent space to a low-rank variety. Then, $\|\mathcal{P}_{(T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))^\perp}(L)\|_2 \leq \|\mathcal{P}_{T'^{\perp}}(L)\|_2$ for any matrix $L\in\mathbb{R}^{p\times p}$.*

Proof [Proof of Lemma 20] Let $\mathcal{R}',\mathcal{C}'$ be row/column space pair that form the tangent space T'. Let $\tilde{C}=\mathrm{span}(\mathcal{C}',\mathbf{1})$ and $\tilde{\mathcal{R}}=\mathrm{span}(\mathcal{R}',\mathbf{1})$. Then, it is straightforward to see that $T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ is itself a tangent space formed by column space $\tilde{C}$ and row space $\tilde{R}$. Thus, $\|\mathcal{P}_{(T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))^\perp}(L^\star)\|_2 = \|\mathcal{P}_{\tilde{\mathcal{C}}^\perp}L^\star\mathcal{P}_{\tilde{\mathcal{R}}^\perp}\|_2$. Since $\mathcal{C}'\subseteq\tilde{\mathcal{C}}$, we have that: $\|\mathcal{P}_{\tilde{\mathcal{C}}^\perp}L\mathcal{P}_{\tilde{\mathcal{R}}^\perp}\|_2 \leq \|\mathcal{P}_{\mathcal{C}'^{\perp}}L\mathcal{P}_{\mathcal{R}'^{\perp}}\|_2 = \|\mathcal{P}_{T'^{\perp}}(L)\|_2$. ■

Lemma 21 *Suppose that $\kappa^\star>\omega$. Then, $\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)\cap(T'\oplus T^\star)=\{0\}$ for every tangent space T' with $\rho(T',T^\star)\leq\omega$.*

Proof [Proof of Lemma 21] It suffices to show that $\|\mathcal{P}_{(T'\oplus T^\star)^\perp}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2>0$. Let $\mathcal{C}'$ be the column space associated with the tangent space T' at a symmetric matrix. Note that $T'\oplus T^\star$ is another tangent space with column space $\mathcal{C}'\oplus\mathcal{C}^\star$. Then, $\|\mathcal{P}_{(T'\oplus T^\star)^\perp}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2 = \|\mathcal{P}_{(\mathcal{C}'\oplus\mathcal{C}^\star)^\perp}(\mathbf{1}/\sqrt{p})\|_2^2$. So it suffices to show that $\|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}(\mathbf{1}/\sqrt{p})\|_2<1$. Note additionally that $\|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}(\mathbf{1}/\sqrt{p})\|_2 \leq \|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}\mathcal{P}_{\mathcal{C}^\star}(\mathbf{1}/\sqrt{p})\|_2 + \|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}\mathcal{P}_{\mathcal{C}^{\star\perp}}(\mathbf{1}/\sqrt{p})\|_2 \leq \|\mathcal{P}_{\mathcal{C}^\star}(\mathbf{1}/\sqrt{p})\|_2 + \|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}\mathcal{P}_{\mathcal{C}^{\star\perp}}\|_2$. We have that: $\|\mathcal{P}_{\mathcal{C}^\star}(\mathbf{1}/\sqrt{p})\|_2 = 1-\kappa^\star$. Using Lemma 18, it is straightforward to conclude that $\|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}\mathcal{P}_{\mathcal{C}^{\star\perp}}\|_2 \leq \|\mathcal{P}_{\mathcal{C}'}\mathcal{P}_{\mathcal{C}^{\star\perp}}\|_2 \leq \|\mathcal{P}_{\mathcal{C}'}-\mathcal{P}_{\mathcal{C}^\star}\|_2$. Appealing to Lemma 17 , and putting everything together, we conclude that: $\|\mathcal{P}_{\mathcal{C}'\oplus\mathcal{C}^\star}(\mathbf{1}/\sqrt{p})\|_2 \leq (1-\kappa^\star)+\omega$. As $\kappa^\star>\omega$, we have the desired result. ■

Lemma 22 *Let $Z=T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\|Z\|_2=1$ and $\rho(T',T^\star)\leq\omega$. Then, assuming $\kappa^\star>\omega$, Z can be decomposed uniquely as follows $Z=Z_1+Z_2$ where $Z_1\in T'$, $Z_2\in\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\max\{\|Z_1\|_2,\|Z_2\|_2\}\leq\frac{2\sqrt{5h}}{1-\sqrt{1-(\kappa^\star-\omega)^2}}$.*

Proof [Proof of Lemma 22] The unique decomposition follows from Lemma 21. Since $\omega < 1$, we have that T' and $T^\star$ have the same dimension. Since $Z_1 \in T' \oplus T^\star$, then $\text{rank}(Z_1) \leq 4h$ (this follows from noting that every matrix inside T' or $T^\star$ has rank at most $2h$ and rank of a sum of matrices is less than the sum of the ranks). Further, $\text{rank}(Z_2) \leq 1$, so that $\text{rank}(Z) \leq 5h$. Therefore, $\|Z\|_F \leq \sqrt{5h}$. Notice that: $\|Z\|_F^2 = \|Z_1 + \mathcal{P}_{T'}(Z_2) + \mathcal{P}_{T'^\perp}(Z_2)\|_F^2 = \|Z_1 + \mathcal{P}_{T'}(Z_2)\|_F^2 + \|\mathcal{P}_{T'^\perp}(Z_2)\|_F^2$. Thus, $\|Z_1 + \mathcal{P}_{T'}(Z_2)\|_F \leq \sqrt{5h}$. Using reverse triangle inequality, we conclude that $\|Z_1\|_F \leq \sqrt{5h} + \|\mathcal{P}_{T'}(Z_2)\|_F$. Now notice that: $\|Z_2\|_F^2 = \|\mathcal{P}_{T'^\perp}(Z_2)\|_F^2 + \|\mathcal{P}_{T'}(Z_2)\|_F^2$, so that: $\sqrt{\|Z_2\|_F^2 - \|\mathcal{P}_{T'^\perp}(Z_2)\|_F^2} = \|\mathcal{P}_{T'}(Z_2)\|_F$. Since Z_2 is rank-1, we have then that: $\|\mathcal{P}_{T'}(Z_2)\|_F = \|Z_2\|_2\sqrt{1 - \|\mathcal{P}_{T'^\perp}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2^2}$. Combining things, we conclude that $\|Z_1\|_F \leq \sqrt{5h} + \|Z_2\|_2\sqrt{1 - \|\mathcal{P}_{T'^\perp}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2^2}$. Notice that $\|\mathcal{P}_{T'^\perp}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2 \geq \|\mathcal{P}_{T^{\star\perp}}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2 - \omega = \kappa^\star - \omega$. Reverse triangle inequality also gives $\|Z_2\|_F \leq \|Z_1\|_F + \sqrt{5h}$. Putting the last bounds together, we have that: $\|Z_2\|_F \leq \frac{2\sqrt{5h}}{1-\sqrt{1-(\kappa^{\star 2}-\omega)^2}}$. Plugging this into a previous bound, we also find that $\|Z_1\|_F \leq \frac{2\sqrt{5h}}{1-\sqrt{1-(\kappa^{\star 2}-\omega)^2}}$. ■

Lemma 23 *Let $T' \subseteq \mathbb{R}^{p\times p}$ be a tangent space to a low-rank variety. Then:* $\max_{N\in T'\oplus \text{span}(\mathbf{1}_p\mathbf{1}_p^\top), \|N\|_2=1} \|N\|_\infty \leq 3\xi(T')$.

Proof [Proof of Lemma 23] Let $(\mathcal{R}', \mathcal{C}')$ be the row/column space pair associated with T'. Let $\tilde{C} = \mathcal{C}' \oplus \text{span}(\mathbf{1}_p)$ and $\tilde{R} = \mathcal{R}' \oplus \text{span}(\mathbf{1}_p)$. Since $Z \in T' \oplus \text{span}(\mathbf{1})$, it is straightforward to show that $Z = \mathcal{P}_{\tilde{\mathcal{C}}} Z \mathcal{P}_{\tilde{\mathcal{R}}^\perp} + Z\mathcal{P}_{\tilde{\mathcal{R}}}$. Therefore, $\|Z\|_\infty \leq \max_i \|\mathcal{P}_{\tilde{\mathcal{C}}}(e_i)\|_2 + \max_i \|\mathcal{P}_{\tilde{\mathcal{R}}}(e_i)\|_2$. Letting $\mathcal{C}_1 = \mathcal{C}'$ and $\mathcal{C}_2 = \text{span}(\mathbf{1}_p)$, and appealing to Lemma 18, we have that:

$$\max_i \|\mathcal{P}_{\tilde{\mathcal{C}}}(e_i)\|_2 \leq 2\max_i \max_{u_1\in\text{span}(\mathbf{1}), \|u_1\|_2=1} 2|u_1^T e_i| + \max_i \max_{u_2\in\mathcal{C}', \|u_2\|_2=1} |u_2^T e_i| \leq 2/\sqrt{p} + \mu[\mathcal{C}'].$$

Analogously, letting $\mathcal{C}_1 = \mathcal{R}'$ and $\mathcal{C}_2 = \text{span}(\mathbf{1})$, and appealing to Lemma 18, we have that:

$$\max_i \|\mathcal{P}_{\tilde{\mathcal{R}}}(e_i)\|_2 \leq 2/\sqrt{p} + \mu[\mathcal{R}'].$$

Since $\xi(T') \geq \max\{[\mu[\mathcal{C}'], \mu[\mathcal{R}']\}$ and $2\xi(T') \geq \frac{2}{\sqrt{p}}$, we conclude the desired result. ■

Lemma 24 *Let T' be a tangent space to the low-rank matrix variety with $\rho(T', T^\star) \leq \omega$ for some $\omega \in (0,1)$. Let $\mathbb{H}' = \Omega^\star \times T'$ and $\mathbb{Q}' = \Omega^\star \times (T' + \text{span}(\mathbf{1}_p\mathbf{1}_p^\top))$. Then for any matrix $N \in \mathbb{R}^{p\times p}$, we have that $|\|\mathcal{P}_{\mathbb{H}'}(N)\|_2 - \|\mathcal{P}_{\mathbb{Q}'}(N)\|_2| \leq 2(\kappa^\star + \omega)$ and $|\|\mathcal{P}_{\mathbb{H}'^\perp}(N)\|_2 - \|\mathcal{P}_{\mathbb{Q}'^\perp}(N)\|_2| \leq 2(\kappa^\star + \omega)$.*

Proof Decompose $N = N_1 + N_2$ where $N_1 \in \mathbb{Q}'$ and $N_2 \in \mathbb{Q}'^\perp$. Thus, $\mathcal{P}_{\mathbb{Q}'}(N) = N_1$. Furthermore, since $\mathbb{H}' \subseteq \mathbb{Q}'$, $\mathcal{P}_{\mathbb{H}'}(N) = \mathcal{P}_{\mathbb{H}'}(N_1)$. From Lemma 19, we have that $\|\mathcal{P}_{\mathbb{H}'}(N_1)\|_2 \geq$

$\|N_1\|_2(1-2(\kappa^\star+\omega))$. Thus, $|\|\mathcal{P}_{\mathbb{Q}'}(N)\|_2 - \|\mathcal{P}_{\mathbb{H}'}(N)\|_2| \leq 2(\kappa^\star+\omega)$. Since for any tangent space to a low-rank variety $F \subseteq \mathbb{R}^{p\times p}$, $\|\mathcal{P}_{F^\perp}(N)\|_2 = \|N\|_2 - \|\mathcal{P}_F(N)\|_2$, we can also conclude that $|\|\mathcal{P}_{\mathbb{H}'^\perp}(N)\|_2 - \|\mathcal{P}_{\mathbb{Q}'^\perp}(N)\|_2| \leq 2(\kappa^\star+\omega)$. ■

Lemma 25 *Let $\mathbb{H}' = \Omega^\star \times T'$ where $\rho(T', T^\star) \leq \omega$. Let $\kappa^\star = \|\mathcal{P}_{T^{\star\perp}}(1/p\mathbf{1}_p\mathbf{1}_p^\top)\|_2$. Suppose that $\min_{\mathbb{Q}'\in U(\omega)} \chi(\mathbb{Q}', \|\cdot\|_{\Phi_\gamma}) > 2(\kappa^\star+\omega)$. Let $F = \mathcal{P}_{T^{\star\perp}}(1/p\mathbf{1}_p\mathbf{1}_p^\top)/\|\mathcal{P}_{T^{\star\perp}}(1/p\mathbf{1}_p\mathbf{1}_p^\top)\|_2$. Then, we have the following results:*

$$\min_{\substack{Z\in\mathbb{H}'\\ \|Z\|_{\Phi_\gamma}=1\\ \rho(T',T^\star)\leq\omega}} \|\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'}(Z)\|_{\Phi_\gamma} \geq \min_{\mathbb{Q}'\in U(\omega)} \chi(\mathbb{Q}', \|\cdot\|_{\Phi_\gamma}) - 2(\kappa^\star+\omega),$$

$$\max_{\substack{Z\in\mathbb{H}'\\ \|Z\|_{\Phi_\gamma}=1\\ \mathbb{Q}'\in U(\omega)}} \|\mathcal{P}_{\mathbb{H}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z)\|_{\Phi_\gamma} \leq \max_{\mathbb{Q}'\in U(\omega)} \varphi(\mathbb{Q}', \|\cdot\|_{\Phi_\gamma}) + 2(\kappa^\star+\omega),$$

$$\max_{\substack{Z\in\mathbb{Q}'\\ \|Z\|_{\Phi_\gamma}=1\\ \mathbb{Q}'\in U(\omega)}} \|(\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'})^{-1}\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'^\perp}(Z)\|_{\Phi_\gamma} \leq \frac{4(\kappa^\star+\omega)\max\{\gamma,1\}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega)}{\min_{\mathbb{Q}'\in U(\omega)} \chi(\mathbb{Q}', \|\cdot\|_{\Phi_\gamma}) - 2(\kappa^\star+\omega)}.$$

where the linear operators $\mathcal{A}, \mathcal{A}^\dagger$, the norm Φ_γ, the set $U(\omega)$, and the functions χ and φ are defined in Section 4.

Proof To prove the first part, consider $Z \in \mathbb{H}'$. Thus, $\mathcal{P}_{\mathbb{Q}'}(Z) = Z$. Combining this with Lemma 24 and noting that the first components of $\mathbb{H}'$ and $\mathbb{Q}'$ are identical, we find that $|\|\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'}(Z)\|_{\Phi_\gamma} - \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma}| \leq 2(\kappa^\star+\omega)$. This results allows us to conclude the first paper.

To prove the second part, let $Z \in \mathbb{H}'$ with $\|Z\|_{\Phi_\gamma} \leq 1$. We first notice that by appealing to Lemma 24, we have that: $\|\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} \geq \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} - 2(\kappa^\star+\omega) > 0$. Thus, the operator $\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}$ is invertible. Furthermore, suppose that there exists $N_1 \in \mathbb{Q}', N_2 \in \mathbb{Q}'$ such that $\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(N_1) = \mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(N_2)$. Since $\mathbb{H}' \subseteq \mathbb{Q}'$, we have that then: $\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(N_1 - N_2) = 0$, which allows us to conclude that for any $Z \in \mathbb{H}'$, $(\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z) = (\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z)$. Appealing to Lemma 24, we have that for any $N \in \mathbb{Q}'$, $|\mathcal{P}_{\mathbb{H}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(N) - \mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(N)| \leq 2(\kappa^\star+\omega)$, which allows us to conclude the desired result.

To prove the third part, Consider any $Z \in \mathbb{Q}'$ with Z_2 denoting its second component which is contained in $T' \oplus \text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$. Let $Z_2 = Z_{21} + Z_{22}$ where $Z_{21} \in T'$ and $Z_{22} \in \text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$. Notice that $\mathcal{P}_{H'^\perp}(Z) = \mathcal{P}_{T'^\perp}(Z_{22})$. By Lemma 19, $\|Z_{22}\|_2 \leq \kappa^\star + \omega$. Furthermore, $\mathcal{P}_{T'^\perp}(Z_{22}) = (\mathcal{P}_{T'^\perp} - \mathcal{P}_{T^{\star\perp}})(Z_{22}) + \mathcal{P}_{T^{\star\perp}})(Z_{22})$. Thus, using the fact that $\|\mathcal{P}_{T'}(M)\|_2 \leq 2\|M\|_2$ for any matrix M, we have that: $\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'^\perp}(Z)\|_{\Phi_\gamma} \leq 4(\kappa^\star + \omega)\max\{\gamma,1\}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega)$. Then, appealing to the first part of the Lemma, we have the desired result.

■

Lemma 26 *Let* $\mathbb{Q}^\star = \Omega^\star \times (T^\star + \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ *and* $\mathbb{Q}' = \Omega^\star \times (T' + \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ *where* $\rho(T', T^\star) \leq \omega$). *Then, for any* $Z \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$ *with* $\|Z\|_{\Phi_\gamma} = 1$,

$$|\|\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} - \|\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}| \leq 5\omega + 4\kappa^\star.$$

Proof *Notice that:*

$$\|\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} - \|(\mathcal{P}_{\mathbb{Q}'}(Z) - \mathcal{P}_{\mathbb{Q}^\star})(Z)\|_{\Phi_\gamma} \leq \|\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} \leq \|\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} + \|(\mathcal{P}_{\mathbb{Q}'}(Z) - \mathcal{P}_{\mathbb{Q}^\star})(Z)\|_{\Phi_\gamma}.$$

Further, letting $Z = (Z_1, Z_2)$ *with* $\|Z_2\|_2/\gamma \leq 1$*:*

$$\begin{aligned}\|(\mathcal{P}_{\mathbb{Q}'} - \mathcal{P}_{\mathbb{Q}^\star})(Z)\|_{\Phi_\gamma} &= \frac{1}{\gamma}\|(\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)} - \mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)})(Z_2)\|_2 \\ &\leq 4(\kappa^\star + \omega) + \frac{1}{\gamma}\|(\mathcal{P}_{T'} - \mathcal{P}_{T^\star})(Z_2)\|_2 \leq 4\kappa^\star + 5\omega.\end{aligned}$$

Combining the results proves our result. ■

Lemma 27 *Let* $\mathbb{Q}^\star = \Omega^\star \times (T^\star + \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ *and* $\mathbb{Q}' = \Omega^\star \times (T' + \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ *where* $\rho(T', T^\star) \leq \omega$). *Suppose*

$$\min_{Z\in\mathbb{Q}^\star, \|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} \geq \beta.$$

Then,

$$\begin{aligned}\min_{Z\in\mathbb{Q}', \|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} \geq \beta(1 - (4\kappa^\star + 5\omega)) \\ - 2(5\omega + 4\kappa^\star)\|\mathbb{I}^\star\|_2 \max\{\gamma, 1\} \\ - (4\kappa^\star + 5\omega)\|\mathbb{I}^\star\|_2(d^\star/\gamma + 1).\end{aligned}$$

Additionally, for any $Z \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$ *with* $\|Z\|_{\Phi_\gamma} = 1$*:*

$$\begin{aligned}|\|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} - \|\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}| \leq 2(5\omega + 4\kappa^\star)\|\mathbb{I}^\star\|_2 \max\{\gamma, 1\} \\ + (4\kappa^\star + 5\omega)\|\mathbb{I}^\star\|_2(d^\star/\gamma + 1).\end{aligned}$$

Proof Consider any Z with $\|Z\|_{\Phi_\gamma} = 1$. Then,

$$\begin{aligned}\|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} &\geq \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} - \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(\mathcal{P}_{\mathbb{Q}'} - \mathcal{P}_{\mathbb{Q}^\star})(Z)\|_{\Phi_\gamma} \\ &\geq \|\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} - \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(\mathcal{P}_{\mathbb{Q}'} - \mathcal{P}_{\mathbb{Q}^\star})(Z)\|_{\Phi_\gamma} \\ &\quad - \|(\mathcal{P}_{\mathbb{Q}'} - \mathcal{P}_{\mathbb{Q}^\star})\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}\end{aligned}$$

Some algebra and appealing to Lemma 26 leads to the conclusions that

$$\|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(\mathcal{P}_{\mathbb{Q}'}-\mathcal{P}_{\mathbb{Q}^\star})(Z)\|_{\Phi_\gamma}\leq 2(5\omega+4\kappa^\star)\|\mathbb{I}^\star\|_2\max\{\gamma,1\},$$

and that $\|\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}\geq(1-(4\kappa^\star+5\omega))$. Furthermore, denote $Z_1=\mathcal{P}_{\Omega^\star}(Z)$ and $Z_2=\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}(Z)$. Notice that $\|Z_1\|_2\leq\|Z_1\|_\infty\theta(\Omega^\star)\leq\|Z_1\|_\infty d^\star$. Again appealing to Lemma 26:

$$\|(\mathcal{P}_{\mathbb{Q}'}-\mathcal{P}_{\mathbb{Q}^\star})\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}\leq(4\kappa^\star+5\omega)\|\mathbb{I}^\star\|_2(d^\star/\gamma+1)$$

Putting things together, we have the first desired result. The second desired result follows from a similar analysis as the first part. ■

Lemma 28 *We begin with the following lemmas where we let* $\mathbb{Q}^\star=\Omega^\star\times(T^\star+\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ *and* $\mathbb{Q}'=\Omega^\star\times(T'+\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))$ *where* $\rho(T',T^\star)\leq\omega$*). Suppose*

$$\min_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1}\|\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}\geq\beta>\frac{2(5\omega+4\kappa^\star)\|\mathbb{I}^\star\|_2\max\{\gamma,1\}}{1-(4\kappa^\star+5\omega)}+\frac{(d^\star+\gamma)(4\kappa^\star+5\omega)\|\mathbb{I}^\star\|_2\max\{\gamma,1\}}{1-(4\kappa^\star+5\omega)},$$

and

$$\max_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1}\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z)\|_{\Phi_\gamma}\leq\zeta,$$

and

$$\max_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1}\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}\leq\delta.$$

Then,

$$\begin{aligned}&\max_{Z\in\mathbb{Q}',\|Z\|_{\Phi_\gamma}=1}\|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z)\|_{\Phi_\gamma}\leq\\&(2\delta+2\|\mathbb{I}^\star\|\max\{\gamma,1\}(4\kappa^\star+5\omega)+\|\mathbb{I}^\star\|(d^\star/\gamma+1)(4\kappa^\star+5\omega))\frac{\|\Delta\|_{\Phi_\gamma}+4\kappa^\star+5\omega}{\beta}\\&+\zeta(1+5\omega+4\kappa^\star)+\|\mathbb{I}^\star\|_2\max\{\gamma,1\}(5\omega+4\kappa^\star)\frac{1+5\omega+4\kappa^\star}{\beta}\\&+\|\mathbb{I}^\star\|_2\max\{\gamma,1\}(5\omega+4\kappa^\star)(1+d^\star/\gamma)\frac{1+5\omega+4\kappa^\star}{\beta},\end{aligned}$$

where,

$$\|\Delta\|_{\Phi_\gamma}\leq\frac{1}{\tilde{\beta}}\left(2(5\omega+4\kappa^\star)\|\mathbb{I}^\star\|_2\max\{\gamma,1\}+(d^\star+\gamma)(4\kappa^\star+5\omega)\|\mathbb{I}^\star\|_2\max\{\gamma,1\}\right),$$

and

$$\tilde{\beta} := \beta - 2(5\omega + 4\kappa^\star)\|\mathbb{I}^\star\|_2 \max\{\gamma, 1\} - (d^\star + \gamma)(4\kappa^\star + 5\omega)\|\mathbb{I}^\star\|_2 \max\{\gamma, 1\}.$$

Proof Take $Z \in \mathbb{Q}'$. Let $Z = \mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(A)$. Define $\Delta := \mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(A) - Z$. Then, by Lemma 27

$$\|\Delta\|_{\Phi_\gamma} \leq \|A\|_{\Phi_\gamma} \left(2(5\omega + 4\kappa^\star)\|\mathbb{I}^\star\|_2 \max\{\gamma, 1\} + (4\kappa^\star + 5\omega)\|\mathbb{I}^\star\|_2(d^\star/\gamma + 1)\right),$$

and that $\|A\|_{\Phi_\gamma} \leq 1/\tilde{\beta}$ where

$$\begin{aligned}\tilde{\beta} := \beta(1 - (4\kappa^\star + 5\omega)) \\ - 2(5\omega + 4\kappa^\star)\|\mathbb{I}^\star\|_2 \max\{\gamma, 1\} \\ - (4\kappa^\star + 5\omega)\|\mathbb{I}^\star\|_2(d^\star/\gamma + 1).\end{aligned}$$

Let $B = (\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}\mathcal{P}_{\mathbb{Q}^\star}(Z)$. Then, $\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(A - B) = \Delta + \mathcal{P}_{\mathbb{Q}^{\star\perp}}(Z)$, which implies that:

$$\|A - B\|_{\Phi_\gamma} \leq \frac{\|\Delta\|_{\Phi_\gamma} + \|\mathcal{P}_{\mathbb{Q}^{\star\perp}}(Z)\|_{\Phi_\gamma}}{\beta} \leq \frac{\|\Delta\|_{\Phi_\gamma} + 4\kappa^\star + 5\omega}{\beta}.$$

Note that:

$$\begin{aligned}&\|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z)\|_{\Phi_\gamma} \leq \\ &\underbrace{\|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}}_{T1} \\ &+ \underbrace{\|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}\left((\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z) - (\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}\mathcal{P}_{\mathbb{Q}^\star}(Z)\right)\|_{\Phi_\gamma}}_{T2}.\end{aligned}$$

Notice that for any $M \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, appealing to Lemma 26

$$\begin{aligned}\|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(M)\|_{\Phi_\gamma} &\leq \|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(M)\|_{\Phi_\gamma}, \\ &+ \|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(\mathcal{P}_{\mathbb{Q}'} - \mathcal{P}_{\mathbb{Q}^\star})(M)\|_{\Phi_\gamma} \\ &+ \|(\mathcal{P}_{\mathbb{Q}'^\perp} - \mathcal{P}_{\mathbb{Q}^{\star\perp}})\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(M)\|_{\Phi_\gamma}, \\ &\leq 2\delta + 2\|\mathbb{I}^\star\| \max\{\gamma, 1\}(4\kappa^\star + 5\omega) + \|\mathbb{I}^\star\|(d^\star/\gamma + 1)(4\kappa^\star + 5\omega).\end{aligned}$$

$$T_2 \leq (2\delta + 2\|\mathbb{I}^\star\| \max\{\gamma, 1\}(4\kappa^\star + 5\omega) + \|\mathbb{I}^\star\|(d^\star/\gamma + 1)(4\kappa^\star + 5\omega))\|A - B\|_{\Phi_\gamma}.$$

To control T_1, notice that for any $M \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, appealing to Lemma 26

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{Q}'^{\perp}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(M)\|_{\Phi_\gamma} &\leq \|\mathcal{P}_{\mathbb{Q}'^{\perp}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma} + \|\mathbb{I}^{\star}\|_2 \max\{\gamma,1\}(5\omega+4\kappa^{\star}) \\
&\leq \|\mathcal{P}_{\mathbb{Q}^{\star}{}^{\perp}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma} + \|(\mathcal{P}_{\mathbb{Q}'^{\perp}} - \mathcal{P}_{\mathbb{Q}^{\star}{}^{\perp}})\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma} \\
&+ \|\mathbb{I}^{\star}\|_2 \max\{\gamma,1\}(5\omega+4\kappa^{\star})\|\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma} \\
&\leq \|\mathcal{P}_{\mathbb{Q}^{\star}{}^{\perp}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma} \\
&+ \|\mathbb{I}^{\star}\|_2 \max\{\gamma,1\}(5\omega+4\kappa^{\star})(1+d^{\star}/\gamma)\|\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma} \\
&+ \|\mathbb{I}^{\star}\|_2 \max\{\gamma,1\}(5\omega+4\kappa^{\star})\|\mathcal{P}_{\mathbb{Q}^{\star}}(M)\|_{\Phi_\gamma}
\end{aligned}$$

Setting $M = (\mathcal{P}_{\mathbb{Q}^{\star}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}})^{-1}\mathcal{P}_{\mathbb{Q}^{\star}}(Z)$ and noting that $\|M\|_{\Phi_\gamma} \leq \frac{1+5\omega+4\kappa^{\star}}{\beta}$, we have the following bound for T_1:

$$\begin{aligned}
T_1 \leq \zeta(1+5\omega+4\kappa^{\star}) &+ \|\mathbb{I}^{\star}\|_2 \max\{\gamma,1\}(5\omega+4\kappa^{\star})\frac{1+5\omega+4\kappa^{\star}}{\beta} \\
&+ \|\mathbb{I}^{\star}\|_2 \max\{\gamma,1\}(5\omega+4\kappa^{\star})(1+d^{\star}/\gamma)\frac{1+5\omega+4\kappa^{\star}}{\beta}
\end{aligned}$$

Combining the bounds on T_1 and T_2, we have the desired result. ■

Appendix E. A numerical approach to verifying Assumptions 1-3

In our numerical approach, we obtain lower bound for

$$\min_{Z\in\mathbb{Q}^{\star},\|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}^{\star}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(Z)\|_{\Phi_\gamma}, \tag{24}$$

and and an upper bound for

$$\max_{Z\in\mathbb{Q}^{\star},\|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}^{\star}{}^{\perp}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(\mathcal{P}_{\mathbb{Q}^{\star}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}})^{-1}(Z)\|_{\Phi_\gamma}. \tag{25}$$

We then can appeal to Lemmas 27-28 to quantify the quantities in Assumptions 1-3. To evaluate (24), consider $Z = (Z_1, Z_2)$ where $Z_1 \in \Omega^{\star}$ with $\|Z_1\|_\infty = 1$ and $Z_2 \in T^{\star} \oplus \text{span}(\mathbf{1}_p\mathbf{1}_p^{\top})$ with $\|Z_2\|_2 \leq \gamma$. Then,

$$\begin{aligned}
\|\mathcal{P}_{\mathbb{Q}^{\star}}\mathcal{A}^{\dagger}\mathbb{I}^{\star}\mathcal{A}\mathcal{P}_{\mathbb{Q}^{\star}}(Z)\|_{\Phi_\gamma} &\geq \|\mathcal{P}_{\Omega^{\star}}\mathbb{I}^{\star}(Z_1)\|_\infty - \|\mathcal{P}_{\Omega^{\star}}\mathbb{I}^{\star}(Z_2)\|_\infty \\
&\geq \min_{Z_1\in\Omega^{\star},\|Z_1\|_\infty=1} \|\mathcal{P}_{\Omega^{\star}}\mathbb{I}^{\star}(Z_1)\|_\infty - \gamma \max_{Z_2\in T^{\star}\oplus\text{span}(\mathbf{1}_p\mathbf{1}_p^{\top}),\|Z_2\|_2=1} \|\mathcal{P}_{\Omega^{\star}}\mathbb{I}^{\star}(Z_2)\|_\infty.
\end{aligned}$$

Now consider $Z = (Z_1, Z_2)$ where $Z_1 \in \Omega^\star$ with $\|Z_1\|_\infty \leq 1$ and $Z_2 \in T^\star \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\|Z_2\|_2 = \gamma$. Then,

$$\begin{aligned}\|\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} &\geq \frac{1}{\gamma}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_2)\|_2 - \frac{1}{\gamma}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_1)\|_2 \\ &\geq \min_{\substack{Z_2\in T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)\\ \|Z_2\|_\infty=1}} \|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_2)\|_2 \\ &\quad - \frac{1}{\gamma}\max_{\substack{Z_1\in\Omega^\star\\ \|Z_1\|_\infty=1}} \|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_1)\|_2.\end{aligned}$$

Thus, we obtain the following lower bound for (24):

$$\begin{aligned}&\min_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} \geq \\ &\min\Bigg\{\min_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1}\|\mathcal{P}_{\Omega^\star}\mathbb{I}^\star(Z_1)\|_\infty - \gamma\max_{Z_2\in T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top),\|Z_2\|_2=1}\|\mathcal{P}_{\Omega^\star}\mathbb{I}^\star(Z_2)\|_\infty, \\ &\quad\min_{\substack{Z_2\in T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)\\ \|Z_2\|_\infty=1}}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_2)\|_2 - \frac{1}{\gamma}\max_{\substack{Z_1\in\Omega^\star\\ \|Z_1\|_\infty=1}}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_1)\|_2\Bigg\}.\end{aligned}$$

The individual terms above are computed approximately by sampling. To obtain an upper bound for (25), consider $Z = (Z_1, Z_2)$ where $Z_1 \in \Omega^\star$ with $\|Z_1\|_\infty = 1$ and $Z_2 \in T^\star \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\|Z_2\|_2 \leq \gamma$. Then,

$$\begin{aligned}&\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z)\|_{\Phi_\gamma} \\ &\leq \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z_1,0)\|_\infty \\ &+ \|\mathcal{P}_{\Omega^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(0,Z_2)\|_\infty \\ &\leq \max_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1}\|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z_1,0)\|_\infty \\ &+ \gamma\max_{Z_2\in T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top),\|Z_2\|=1}\|\mathcal{P}_{\Omega^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(0,Z_2)\|_\infty\end{aligned}$$

Now consider $Z = (Z_1, Z_2)$ where $Z_1 \in \Omega^\star$ with $\|Z_1\|_\infty$ $e1$ and $Z_2 \in T^\star \oplus \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\|Z_2\|_2 = \gamma$. Then,

$$\begin{aligned}&\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z)\|_{\Phi_\gamma} \\ &\leq \frac{1}{\gamma}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(0,Z_2)\|_2 \\ &+ \frac{1}{\gamma}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z_1,0)\|_2 \\ &\leq \max_{Z_2\in T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top),\|Z_2\|_2=1}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(0,Z_2)\|_2 \\ &+ \frac{1}{\gamma}\max_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1}\|\mathcal{P}_{T^\star\oplus\operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z_1,0)\|_2\end{aligned}$$

Thus, we obtain the following upper bound for (25):

$$
\begin{aligned}
&\max_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z)\|_{\Phi_\gamma} \\
&\leq \max\Bigg\{ \max_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1} \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z_1,0)\|_\infty \\
&+\gamma \max_{Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top),\|Z_2\|=1} \|\mathcal{P}_{\Omega^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(0,Z_2)\|_\infty \\
&, \max_{Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top),\|Z_2\|_2=1} \|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(0,Z_2)\|_2 \\
&+\frac{1}{\gamma}\max_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1} \|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(\mathcal{P}_{\mathbb{Q}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star})^{-1}(Z_1,0)\|_2 \Bigg\}
\end{aligned}
$$

Again, the individual terms above are computed approximately by sampling. Finally, we note that appealing to Lemmas 28 involves computing an upper-bound for:

$$
\max_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1} \|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}. \tag{26}
$$

To obtain an upper bound, consider $Z=(Z_1,Z_2)$ where $Z_1\in\Omega^\star$ with $\|Z_1\|_\infty=1$ and $Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\|Z_2\|_2\leq\gamma$. Then,

$$
\begin{aligned}
\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} &\leq \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_1)\|_\infty + \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_2)\|_\infty \\
&\leq \max_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1} \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_1)\|_\infty + \gamma \max_{\substack{Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)\\ \|Z_2\|_2=1}} \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_2)\|_\infty.
\end{aligned}
$$

Now consider $Z=(Z_1,Z_2)$ where $Z_1\in\Omega^\star$ with $\|Z_1\|_\infty\leq 1$ and $Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ with $\|Z_2\|_2=\gamma$. Then,

$$
\begin{aligned}
\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma} &\leq \frac{1}{\gamma}\|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_2)\|_2 + \frac{1}{\gamma}\|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_1)\|_2 \\
&\geq \max_{\substack{Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)\\ \|Z_2\|_\infty=1}} \|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_2)\|_2 \\
&\quad + \frac{1}{\gamma}\max_{\substack{Z_1\in\Omega^\star\\ \|Z_1\|_\infty=1}} \|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_1)\|_2.
\end{aligned}
$$

Thus, we obtain the following lower bound for (26):

$$\max_{Z\in\mathbb{Q}^\star,\|Z\|_{\Phi_\gamma}=1}\|\mathcal{P}_{\mathbb{Q}^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}^\star}(Z)\|_{\Phi_\gamma}\geq$$

$$\max\Bigg\{\max_{Z_1\in\Omega^\star,\|Z_1\|_\infty=1}\|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_1)\|_\infty+\gamma\max_{\substack{Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)\\ \|Z_2\|_2=1}}\|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_2)\|_\infty,$$

$$\max_{\substack{Z_2\in T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)\\ \|Z_2\|_\infty=1}}\|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_2)\|_2+\frac{1}{\gamma}\max_{\substack{Z_1\in\Omega^\star\\ \|Z_1\|_\infty=1}}\|\mathcal{P}_{T^\star\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_1)\|_2\Bigg\}.$$

Appendix F. Sufficient Hessian conditions and choice of γ that satisfies Assumptions 1-3

Behavior of $\mathbb{I}^\star$ with respect to $\Omega^\star$. Let

$$\begin{aligned}\alpha_\Omega &:= \min_{N\in\Omega^\star,\|N\|_\infty=1}\|\mathcal{P}_{\Omega^\star}\mathbb{I}^\star\mathcal{P}_{\Omega^\star}(N)\|_\infty,\\ \delta_{\Omega^\perp}&: \max_{N\in\Omega^\star,\|N\|_\infty=1}\|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star\mathcal{P}_{\Omega^\star}(N)\|_\infty,\\ \beta_\Omega &:= \max_{N\in\Omega^\star,\|N\|_2=1}\|\mathbb{I}^\star(N)\|_2,\end{aligned}$$

be functions $\mathbb{I}^\star$ with respect to $\Omega^\star$. Here, α_Ω quantifies the minimum gain of $\mathbb{I}^\star$ restricted to subspace $\Omega^\star$ and with respect to the ℓ_∞ norm (the minimum gain of a matrix M restricted to subspace S and with respect to norm $\|\cdot\|$ is $\min_{x\in S,\|x\|=1}\|P_S M P_S(x)\|$); the quantity δ_Ω computes the inner-product between elements in $\Omega^\star$ and $\Omega^{\star\perp}$ as quantified by the metric induced by $\mathbb{I}^\star$; and finally, β_Ω quantifies the behavior of $\mathbb{I}^\star$ restricted to $\Omega^\star$ in spectral norm.

Behavior of $\mathbb{I}^\star$ with respect to $T^\star$. Similar to $\Omega^\star$, we control the behavior of $\mathbb{I}^\star$ associated with the subspace $T^\star$. We control the behavior of $\mathbb{I}^\star$ for tangent spaces T' close to the tangent space $T^\star$:

$$\begin{aligned}\alpha_T &:= \min_{N\in T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top),\rho(T',T^\star)\leq\omega,\|N\|_2=1}\|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}(N)\|_2,\\ \delta_{T^\perp} &:= \max_{\substack{N\in T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top),\rho(T',T^\star)\leq\omega,\\ \|N\|_2\leq1}}\|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}(N)\|_2,\\ \beta_T &:= \max_{N\in T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top),\rho(T',T^\star)\leq\omega,\|N\|_\infty=1}\|\mathbb{I}^\star(N)\|_\infty.\end{aligned}$$

Here, α_T quantifies the minimum gain of $\mathbb{I}^\star$ restricted to tangent spaces $T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ that are close to $T^\star$ with respect to the spectral norm; the quantify δ_T computes the inner-product between elements in T' and $T'^\perp$ as quantified by the metric induced by $\mathbb{I}^\star$; and finally, β_T quantifies the behavior of $\mathbb{I}^\star$ restricted to $T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ and $T'^\perp$ in infinity norm.

With these above quantities defined, and letting $\tilde{\alpha}:=\min\{\alpha_\Omega,\alpha_T\}$, $\tilde{\delta}:=\max\{\delta_{\Omega^\perp},\delta_{T^\perp}\}$, and $\tilde{\beta}:=\max\{\beta_\Omega,\beta_T\}$, the main assumptions are the following. Recall that $d^\star:=\max_i\sum_{j=1}^p\mathbb{I}[|S^\star_{ij}|>0$

represents the maximal degree of the conditional graphical structure of the observed variables conditioned on the latent variables and $\mu^\star := \max_i \|\mathcal{P}_{\text{col-space}(L^\star)} e_i\|_2$ represents the denseness of the latent effects with e_i denoting a standard coordinate basis element.

Assumption 4 $\tilde{\alpha} > 0$.

Assumption 5 *There exists* $\tilde{\nu} \in (2\omega, 1/2)$ *such that* $\tilde{\delta}/\tilde{\alpha} \leq 1 - 2\tilde{\nu}$.

Assumption 6 *The product of degree of sparsity of* $S^\star$, $d^\star$, *and the diffuseness of the latent effects,* $\mu^\star$, *is bounded as follows:* $d^\star(6\mu^\star + \omega) \leq \frac{\tilde{\nu}^2\tilde{\alpha}^2}{2\tilde{\beta}^2(2-\tilde{\nu})^2}$ *where* $\tilde{\nu}$ *and* $d^\star \leq \frac{\tilde{\alpha}^2\tilde{\nu}}{32\omega\tilde{\beta}(2-\tilde{\nu}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega)}$.

Assumption 7 *The regularization parameter* γ *chosen in the following range:*

$$\gamma \in \left[\frac{2\tilde{\beta}d^\star(2-\tilde{\nu})}{\tilde{\nu}\tilde{\alpha}}, \min\left\{\frac{\tilde{\nu}\tilde{\alpha}(1-\omega)}{\tilde{\beta}(6\mu^\star+\omega)(2-\tilde{\nu})}, \frac{\tilde{\alpha}}{16\omega(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega + 1)}\right\}\right]$$

Assumption 8 $\kappa^\star := \|\mathcal{P}_{T^{\star\perp}}(\mathbf{1}_p\mathbf{1}_p^\top/p)\|_2 \in \left(\omega, \min\left\{2\tilde{\nu}, \frac{\tilde{\alpha}}{16\max\{\gamma,1\}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega + 1)} - \omega\right\}\right)$.

Assumptions (4)-6 are akin to conditions imposed in Chandrasekaran et al. (2012), although our conditions the subspace $T' \oplus \text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ that arises from the additional zero row sum constraint in our estimator. Assumption 6 ensures that $d^\star$ and $\mu^\star$ are not simultaneously large, and this type of condition was shown to be sufficient for recovering a sparse and low-rank matrix from their sum using mix of ℓ_1 and nuclear norm regularization (Chandrasekaran et al., 2011). Assumption 8 is a new condition relative to Chandrasekaran et al. (2012) to deal with the dual parameter $t\mathbf{1}_p\mathbf{1}_p^\top$ that arises from the zero sum constraint.

Lemma 29 *Under Assumptions 5-8, we have that Hessian assumptions 1-3 for some* $\alpha = \tilde{\alpha}/2$, $\nu = 2\tilde{\nu}$.

Proof Our analysis will depend on the following quantities for any pair of subspaces $\Omega, T \subseteq \mathbb{R}^{p\times p}$:

$$\theta(\Omega) := \max_{N\in\Omega, \|N\|_\infty = 1} \|N\|_2 \quad ; \quad \xi(T) := \max_{N\in T, \|N\|_2 = 1} \|N\|_\infty.$$

When $\Omega = \Omega^\star$ and $T = T^\star$, these quantities are closely connected to the maximal degree $d^\star$ and the incoherence parameter $\mu^\star$ (defined in Section 4.1). In particular, Chandrasekaran et al. (2012) showed that $\mu(\Omega^\star) \in [0, d^\star]$ and $\xi(T^\star) \in [\mu^\star, 2\mu^\star]$.

We consider the quantity $\min_{Z\in\mathbb{Q}', \Phi_\gamma(Z)=1} \|\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma}$. Let $Z = (Z_1, Z_2)$ where $\|Z\|_{\Phi_\gamma} = 1$, $Z \in \mathbb{Q}'$. Suppose $\|Z_1\|_\infty = 1$. Then using Lemmas 16 and 23, we have that:

$$\begin{aligned}
\|\mathcal{P}_{\Omega^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(Z)\|_\infty &\geq \|\mathcal{P}_{\Omega^\star}\mathbb{I}^\star\mathcal{P}_{\Omega^\star}(Z_1)\|_\infty - \|\mathcal{P}_{\Omega^\star}\mathbb{I}^\star(Z_2)\|_\infty \\
&\geq \alpha' - \|\mathbb{I}^\star(Z_2)\|_\infty \\
&\geq \tilde{\alpha} - \gamma\tilde{\beta}\xi(T' \oplus \text{span}(\mathbf{1}_p\mathbf{1}_p^\top)) \geq \tilde{\alpha} - 3\gamma\tilde{\beta}\xi(T') \geq \tilde{\alpha} - \frac{(3\xi(T^\star)+\omega)}{1-\omega}\tilde{\beta}\gamma \\
&\geq \tilde{\alpha} - \frac{\tilde{\nu}\tilde{\alpha}}{2-\tilde{\nu}}.
\end{aligned}$$

Now suppose that $\|Z_2\|_2 = \gamma$. Then, we have using Lemma 19 that:

$$\begin{aligned}\|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star\mathcal{A}(Z)\|_2 &\geq \|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}(Z_2)\|_2 \\ &\quad - \|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}\mathbb{I}^\star(Z_1)\|_2 \\ &\geq \tilde{\alpha}\gamma - 2\tilde{\beta}\theta(\Omega^\star) \geq \tilde{\alpha}\gamma - \frac{\tilde{\nu}\tilde{\alpha}\gamma}{2-\tilde{\nu}}.\end{aligned}$$

Putting the previous bounds together, we have that:

$$\min_{Z\in\mathbb{Q}',\Phi_\gamma(Z)=1} \|\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} \geq \tilde{\alpha} - \frac{\tilde{\nu}\tilde{\alpha}}{2-\tilde{\nu}} \geq \tilde{\alpha}/2. \tag{27}$$

Now we consider the quantity $\max_{Z\in\mathbb{Q}',\Phi_\gamma(Z)=1} \|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma}$. Let $Z = (Z_1, Z_2)$ where $\|Z\|_{\Phi_\gamma} = 1$, $Z \in \mathbb{Q}'$. Suppose $\|Z_1\|_\infty = 1$.

$$\begin{aligned}\|\mathcal{P}_{\Omega^{\star\perp}}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(Z)\|_\infty &\leq \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star\mathcal{P}_{\Omega^\star}(Z_1)\|_\infty + \|\mathcal{P}_{\Omega^{\star\perp}}\mathbb{I}^\star(Z_2)\|_\infty \\ &\leq \tilde{\delta} + \|\mathbb{I}^\star(Z_2)\|_\infty \leq \tilde{\delta} + \frac{\tilde{\beta}\gamma(3\xi(T^\star)+\omega)}{1-\omega} \leq \tilde{\delta} + \frac{\tilde{\nu}\tilde{\alpha}}{(2-\tilde{\nu})}.\end{aligned}$$

Now suppose that $\|Z_2\|_2 = \gamma$. Then, we have using Lemma 19:

$$\begin{aligned}\|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star\mathcal{A}(Z)\|_2 &\leq \|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)}(Z_2)\|_2 + \|\mathcal{P}_{T'\oplus\mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)^\perp}\mathbb{I}^\star(Z_1)\|_2 \\ &\leq \tilde{\delta}\gamma + \tilde{\beta}\theta(\Omega^\star) \leq \tilde{\delta}\gamma + \frac{\tilde{\nu}\tilde{\alpha}\gamma}{(2-\tilde{\nu})}.\end{aligned}$$

Combining the last two inequalities, we have:

$$\max_{Z\in\mathbb{Q}',\Phi_\gamma(Z)=1} \|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(Z)\|_{\Phi_\gamma} \leq \tilde{\delta} + \frac{\tilde{\nu}\tilde{\alpha}}{(2-\tilde{\nu})}. \tag{28}$$

Combining (27) and (28), we have that:

$$\max_{\substack{Z\in\mathbb{Q}'\\ \|Z\|_\Psi=1}} \|\mathcal{P}_{\mathbb{Q}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(\mathcal{P}_{\mathbb{Q}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z)\|_\Psi \leq \frac{\tilde{\delta} + \frac{\tilde{\nu}\tilde{\alpha}}{(2-\tilde{\nu})}}{\tilde{\alpha} - \frac{\tilde{\nu}\tilde{\alpha}}{2-\tilde{\nu}}} \leq 1-\tilde{\nu}.$$

■

Appendix G. Finite sample convergence guarantees of the empirical variogram matrix

In addition to the identifiability assumptions, following Engelke et al. (2022c), we impose conditions to characterize the convergence rate of the empirical variogram matrix to the population variogram matrix. Throughout, we suppose that the random vector $X = (X_O, X_H)$ is in the domain of attraction of the multivariate Pareto distribution Y following a latent Hüsler–Reiss distribution with parameter matrix Γ; for details see Section 2.1 and 3.1.

Assumption 9 *The marginal distribution functions F_i of X_i, $i \in O$, are continuous and there exists constants $\xi > 0$, $K < \infty$ such that for all triples of distinct indices $J = (i, j, m) \subset O$ and $q \in (0, 1]$,*

$$\sup_{x \in [0, q^{-1}]^2 \times [0,1]} \left| q^{-1} \mathbb{P}(F_J(X_J) > 1 - qx) - \frac{\mathbb{P}(Y_J > 1/x)}{\mathbb{P}(Y_1 > 1)} \right| \leq K q^{\xi},$$

where $F_J(x) = (F_i(x_i), F_j(x_j), F_m(x_m))$.

Assumption 9 is a second-order condition that essentially controls the speed of convergence of the sample variogram matrix to the population variogram matrix.

Corollary 30 *(Engelke et al., 2022c, Theorem 1) Let Assumption 9 hold. Let $\ell \in (0, 1]$ be arbitrary. Suppose that $n^{\ell} \leq k \leq n/2$ where k is the effective sample size in computing the sample variogram matrix (see Section 3.3.1). Let $\vartheta \geq 0$ be any scalar satisfying $\vartheta \leq \sqrt{k}/(\log n)^4$. Then, there exists positive constants $c_5, C_5, \tilde{C}_5$ only depending on K, ξ, ℓ, ϵ, and $G(z)$ such that:*

$$\mathbb{P}\left(\|\hat{\Gamma}_O - \Gamma^{\star}_O\|_{\infty} > C_5 \left\{ \left(\frac{k}{n}\right)^{\xi} (\log(n/k))^2 + \frac{1 + \vartheta}{\sqrt{k}} \right\} \right) \leq \tilde{C}_5 p^3 e^{-c_5 \vartheta^2}.$$

Further, if the random vector X is in the domain of attraction of a max-stable distribution, then $\xi = 1$.

Appendix H. Proof of Theorem 9

H.1 Implied Hessian conditions

Combining Lemma 25 with Assumptions 1-3, and letting $m = \max\{\gamma, 1\}$, we have that the following three properties:

$$\min_{\substack{Z \in \mathbb{H}' \\ \|Z\|_{\Phi_\gamma} = 1 \\ \rho(T', T^{\star}) \leq \omega}} \|\mathcal{P}_{\mathbb{H}'} \mathcal{A}^{\dagger} \mathbb{I}^{\star} \mathcal{A} \mathcal{P}_{\mathbb{H}'}(Z)\|_{\Phi_\gamma} \geq \alpha - 2(\kappa^{\star} + \omega) \in (0, \infty),$$

$$\max_{\substack{Z \in \mathbb{H}' \\ \|Z\|_{\Phi_\gamma} = 1 \\ \mathbb{Q}' \in U(\omega)}} \|\mathcal{P}_{\mathbb{H}'^{\perp}} \mathcal{A}^{\dagger} \mathbb{I}^{\star} \mathcal{A} \mathcal{P}_{\mathbb{Q}'} (\mathcal{P}_{\mathbb{H}'} \mathcal{A}^{\dagger} \mathbb{I}^{\star} \mathcal{A} \mathcal{P}_{\mathbb{Q}'})^{-1}(Z)\|_{\Phi_\gamma} \leq 1 - (\nu - 2(\kappa^{\star} + \omega)) \in [0, 1),$$

$$\max_{\substack{Z \in \mathbb{Q}' \\ \|Z\|_{\Phi_\gamma} = 1 \\ \mathbb{Q}' \in U(\omega)}} \|(\mathcal{P}_{\mathbb{H}'} \mathcal{A}^{\dagger} \mathbb{I}^{\star} \mathcal{A} \mathcal{P}_{\mathbb{H}'})^{-1} \mathcal{P}_{\mathbb{H}'} \mathcal{A}^{\dagger} \mathbb{I}^{\star} \mathcal{A} \mathcal{P}_{\mathbb{H}'^{\perp}}(Z)\|_{\Phi_\gamma} \leq 1 - \frac{4(\kappa^{\star} + \omega) m (\|\mathbb{I}^{\star}(F)\|_2 + \|\mathbb{I}^{\star}\|_2 \omega)}{\alpha - 2(\kappa^{\star} + \omega)} \in [0, 1).$$

The first property follows from $\kappa^{\star} < \frac{\alpha}{4}$ and $\alpha > 4\omega$. The second property follow from $1 - \nu + 2(\kappa^{\star} + \omega) < 1$ since $\nu > 2(\kappa^{\star} + \omega)$. The final property follows from having $\frac{4(\kappa^{\star} + \omega) \max\{\gamma, 1\} (\|\mathbb{I}^{\star}(F)\|_2 + \|\mathbb{I}^{\star}\|_2 \omega + 1)}{\alpha} < 1$ or equivalently that $\kappa^{\star} \leq \frac{\alpha}{8 \max\{\gamma, 1\} (\|\mathbb{I}^{\star}(F)\|_2 + \|\mathbb{I}^{\star}\|_2 \omega + 1)} -$

ω with $\alpha > 8\omega \max\{\gamma, 1\}(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega + 1)$. For notational simplicity and with slight abuse of notation, we let:

$$\alpha' := \alpha - 2(\kappa^\star + \omega),$$
$$\zeta := \max\left\{\frac{1}{\nu - 2(\kappa^\star + \omega)}, \frac{\alpha - 2(\kappa^\star + \omega)}{4(\kappa^\star + \omega)m(\|\mathbb{I}^\star(F)\|_2 + \|\mathbb{I}^\star\|_2\omega)}\right\}.$$

Then, we have the following Hessian conditions:

$$\begin{aligned}
&p1) \quad \min_{\substack{Z\in\mathbb{H}' \\ \|Z\|_{\Phi_\gamma}=1 \\ \rho(T',T^\star)\le\omega}} \|\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'}(Z)\|_{\Phi_\gamma} \ge \alpha' \in (0,\infty), \\
&p2) \quad \max_{\substack{Z\in\mathbb{H}' \\ \|Z\|_{\Phi_\gamma}=1 \\ \mathbb{Q}'\in U(\omega)}} \|\mathcal{P}_{\mathbb{H}'^\perp}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'}(\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{Q}'})^{-1}(Z)\|_{\Phi_\gamma} \le 1 - \frac{1}{\zeta} \in [0,1), \\
&p3) \quad \max_{\substack{Z\in\mathbb{Q}' \\ \|Z\|_{\Phi_\gamma}=1 \\ \mathbb{Q}'\in U(\omega)}} \|(\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'})^{-1}\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'^\perp}(Z)\|_{\Phi_\gamma} \le 1 - \frac{1}{\zeta} \in [0,1),
\end{aligned} \tag{29}$$

H.2 Full theoretical statement

Let $c_5, C_5, \tilde{C}_5$ be constants that ensure Corollary 30 is satisfied. Let $\psi = \max\{1, \|(S^\star - L^\star)^+\|_2\}$, $C_0 = 8 + \frac{32\sqrt{5h}}{\alpha'(1-\sqrt{1-(\kappa^{\star 2}-\omega)^2})(\frac{1}{\zeta}-2(\kappa^\star+\omega))}\left[1 + \frac{1}{3\zeta}\right]$, $C_1 = \psi(m + d^\star)$, and $C_2 = m\max\{\left(\frac{4C_0}{\alpha'} + \frac{1}{\psi}\right), 1\}$. We also define,

$$C_4 = \min\left\{\min\left\{\frac{8\alpha'}{C_1}, \frac{\min\{\alpha', 1\}(\frac{1}{\zeta} - 2(\kappa^\star+\omega))}{16m\psi C_2^2}\right\}\frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star+\omega))}{4(1+\frac{1}{3\zeta})}, \frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star+\omega))}{64C_1(1+\frac{1}{3\zeta})}, \frac{\alpha'^2(\frac{1}{\zeta} - 2(\kappa^\star+\omega))^2}{6144\zeta(1+\frac{1}{3\zeta})^2}\right\}.$$

Theorem 31 *Suppose that there exists $\alpha > 0$, $\nu \in (0, 1]$, $\omega \in (0, 1)$ and the choice of the parameter γ so that the Hessian $\mathbb{I}^\star$ satisfies Assumptions 1-3. Let $m := \max\{1, 1/\gamma\}$ and $\bar{m} := \max\{1, \gamma\}$. Let the effective sample size k be chosen such that $k = o(\lfloor n^{2\xi/(1+2\xi)}\rfloor)$. Furthermore, suppose that:*

$$k \ge \max\left\{\frac{C_5^2 1152 m^2\zeta^2 p^2\log(\tilde{C}_5 p)}{C_4^2 c_5^2} + \frac{72m^2\zeta^2}{C_4^2}, \left(\frac{2C_5}{0.1^2\sqrt{c_5}}\sqrt{\log(\tilde{C}_5 p)}\right)^{-2/(3/2-(2\xi+1)/(2\xi))}, \log(k)^{2/(3/2-(2\xi+1)/(2\xi))}, 4(3/2-(2\xi+1)/(2\xi))^8\frac{\log(\tilde{C}_5 p)}{c_5}\log(k)^8\right\}$$

and

1. $\lambda_n = C_5 \left[\frac{24m\zeta}{\sqrt{c_5}} \sqrt{\frac{p^2 \log(\tilde{C}_5 p)}{k}} + \frac{6m\zeta}{\sqrt{k}} \right]$,

2. $\sigma_{\min}(L^\star) \geq \max \left\{ 16m\bar{m}\frac{\lambda_n C_2}{\omega}, \frac{2\psi C_2^2 \lambda_n}{C_0}, \left(mC_2 + \frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{4\left[1 + \frac{1}{3\zeta}\right]} \right) \lambda_n \right\}$,

3. $|S^\star_{ij}| \geq 12m\bar{m}\lambda_n C_2$ *whenever* $|S^\star_{ij}| > 0$.

Then, the estimate $(\hat{S}, \hat{L})$ *is the unique minimizer of* (9) *with*

$$\mathbb{P}\left(\text{sign}(\hat{S}) = \text{sign}(S^\star), \text{rank}(\hat{L}) = \text{rank}(L^\star), \|(\hat{S} - \hat{L}) - \tilde{\Theta}^\star\|_2 \leq 2mC_2\lambda_n \right) \geq 1 - \frac{1}{p}.$$

To arrive at the scalings provided in Theorem 9, note that, $\zeta = \mathcal{O}(1/\nu)$, $\zeta = \mathcal{O}(1/\nu)$, $C_0 = \mathcal{O}(\sqrt{h}\nu/\alpha')$, $C_1 = \mathcal{O}(md^\star)$, $C_2 = \mathcal{O}(m\sqrt{h}/\alpha'^2)$, $C_4 = \mathcal{O}(\alpha'^3\nu/(d^\star m^3\sqrt{h})$. This scaling allows us to conclude that: $k \gtrsim \frac{m^3 h d^{\star 2}}{\alpha'^6 \nu^2} m\nu p \log(p)$, $\lambda_n = \frac{m}{\nu}\sqrt{\frac{p^2\log(p)}{k}}$, $\sigma_{\min}(L^\star) \gtrsim \frac{m^4 h\bar{m}}{\nu\alpha'^4}\sqrt{\frac{p^2\log(p)}{k}}$, $S^\star_{ij} \gtrsim \frac{m^3\bar{m}\sqrt{h}}{\nu\alpha'^2}\sqrt{\frac{p^2\log(p)}{k}}$, and finally $\|(\hat{S} - \hat{L}) - \tilde{\Theta}^\star\|_2 \lesssim \frac{m^3\sqrt{h}}{\nu\alpha'^2}\sqrt{\frac{p^2\log(p)}{k}}$.

H.3 Proof strategy

The high-level proof strategy is similar in spirit to the proofs of consistency results for sparse graphical model recovery and latent variable graphical model recovery (Chandrasekaran et al., 2012), although our convex program and the conditions required for its success are different from these previous results. Consider the following convex program

$$\begin{aligned} (\hat{S}, \hat{L}) = \arg \min_{S, L \in \mathbb{S}^p} \quad & -\log\det(U^T(S-L)U) - \text{tr}((S-L)\hat{\Gamma}_O/2) + \lambda_n(\|S\|_1 + \gamma\|L\|_\star). \\ \text{subject-to} \quad & S - L \in \text{span}(\mathbf{1}_p\mathbf{1}_p^\top) \end{aligned} \tag{30}$$

Comparing (30) with the convex program (9), the differences are: i) we have removed the positive-definite constraints, ii) we have replaced $\text{tr}(L)$ with $\|L\|_\star$ which is valid for positive semi-definite L, iii) we have replaced the constraint $(S-L)\mathbf{1}_p = 0$ with $S - L \in \text{span}(\mathbf{1}_p\mathbf{1}_p^\top)$ which is equivalent since the matrices S, L are symmetric. Regarding item i), the positive definiteness of $\hat{S} - \hat{L}$ is automatically met due to the log-det term. We show with high probability that $\hat{L} \succeq 0$.

Note that due to the log-det term, we have that $UU^T(S-L)UU^T = S - L$. Appealing to Lemma 13, we conclude that $U(U^T(S-L)U)^{-1}U^T$, which is the gradient of the negative log-determinate term with respect to S is equivalent to $(S-L)^+$. Similarly, since $\text{tr}((S-L)\hat{\Gamma}_0/2) = \text{tr}(UU^T(S-L)UU^T\hat{\Gamma}_0/2) = \text{tr}((S-L)UU^T\hat{\Gamma}_0/2UU^T)$, the gradient of the trace term in the objective with respect to S is given by $UU^T\hat{\Gamma}_0/2UU^T$. Standard convex analysis states that $(\hat{S}, \hat{L})$ is the solution of the convex program (30) if there exists a dual variable

$t \in \mathbb{R}$ with the following conditions being satisfied:

$$\begin{aligned} -UU^T(\hat{\Gamma}_O/2)UU^T - (\hat{S}-\hat{L})^+ + t\mathbf{1}_p\mathbf{1}_p^\top &= -\lambda\partial\|\hat{S}\|_1, \\ UU^T(\hat{\Gamma}_O/2)UU^T + (\hat{S}-\hat{L})^+ - t\mathbf{1}_p\mathbf{1}_p^\top &= -\lambda\gamma\partial\|\hat{L}\|_\star, \\ \hat{S}-\hat{L} &\in \text{span}(\mathbf{1}_p\mathbf{1}_p^\top). \end{aligned} \tag{31}$$

Recall that elements of the subdifferential with respect to nuclear norm at a matrix M have the key property that they decompose with respect to the tangent space $T(M)$. Specifically, the subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = U_l Q U_r^T$ is as follows:

$$N \in \partial\|M\|_\star \Leftrightarrow \mathcal{P}_{T(M)}(N) = U_l V_r^T, \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \leq 1,$$

where $\mathcal{P}$ denotes a projection operator. Similarly, we have the following for the subdifferential of ℓ_1 norm:

$$N \in \partial\|M\|_1 \Leftrightarrow \mathcal{P}_{\Omega(M)}(N) = \text{sign}(N), \|\mathcal{P}_{\Omega(M)^\perp}(N)\|_\infty \leq 1.$$

Let SVD of $\hat{L}$ be $\hat{U}\hat{D}\hat{V}^T$ and let $Z = (-\lambda\text{sign}(\hat{S}), -\lambda\gamma\hat{U}\hat{V}^T)$. Then, letting $\mathbb{H} = \Omega(\hat{S}) \times T(\hat{L})$ the optimality conditions of (30) reduce to:

$$\begin{aligned} \mathcal{P}_{\mathbb{H}}\mathcal{A}^\dagger(-UU^T\hat{\Gamma}_O/2UU^T - (\hat{S}-\hat{L})^+ - t\mathbf{1}_p\mathbf{1}_p^\top) &= Z, \\ \Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger(-UU^T\hat{\Gamma}_O/2UU^T - (\hat{S}-\hat{L})^+ - t\mathbf{1}_p\mathbf{1}_p^\top)) &\leq \lambda_n, \\ \hat{S}-\hat{L} &\in \text{span}(\mathbf{1}_p\mathbf{1}_p^\top). \end{aligned} \tag{32}$$

To ensure that the estimates $(\hat{S}, \hat{L})$ are close to their respective population parameters, the quantity $\Delta_S = \hat{S} - S^\star$ and $\Delta_L = \hat{L} - L^\star$ must be small. Since the optimality conditions of (30) are stated in terms of $(\hat{S}-\hat{L})^+$, we bound the deviation between $(\hat{S}-\hat{L})^+$ and $(S^\star - L^\star)^+$. Specifically, the Taylor Series expansion of $(\hat{S}-\hat{L})^+$ around $(S^\star - L^\star)^+$ is:

$$(\hat{S}-\hat{L})^+ = (S^\star - L^\star + \mathcal{A}(\Delta_S, \Delta_L))^+ = (S^\star - L^\star)^+ + (S^\star - L^\star)^+\mathcal{A}(\Delta_S, \Delta_L)(S^\star - L^\star)^+ + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L).$$

where some algebra yields the following representation for the remainder term $\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\Delta_S, \Delta_L))$:

$$\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\Delta_S, \Delta_L)) = U(S^\star - L^\star + \mathbf{1}_p\mathbf{1}_p^\top/p)^{-1}\left[\sum_{k=2}^{\infty}(-\mathcal{A}(\Delta_S, \Delta_L)(S^\star - L^\star + \mathbf{1}_p\mathbf{1}_p^\top/p)^{-1})^k\right]U^T. \tag{33}$$

From Theorem 5, we have that $(S-L)^+ = UU^T(-\Gamma^\star/2)UU^T$. Since $UU^T(S^\star - L^\star)UU^T = S^\star - L^\star$, we appeal to Lemma 13 to conclude that $(U^T(S^\star - L^\star)U)^{-1} = U^T(-\Gamma_O^\star)U$. Let $E_n := UU^T(\hat{\Gamma}_O - \Gamma^\star)/2UU^T$. Then, we have the following equivalent characterization of the optimality conditions (31):

$$\begin{aligned} &\mathcal{P}_{\mathbb{H}}\mathcal{A}^\dagger((S^\star - L^\star)^+\mathcal{A}(\Delta_S, \Delta_L)(S^\star - L^\star)^+ + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L) + E_n + t\mathbf{1}_p\mathbf{1}_p^\top) = Z, \\ &\Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger((S^\star - L^\star)^+\mathcal{A}(\Delta_S, \Delta_L)(S^\star - L^\star)^+ + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L) + E_n + t\mathbf{1}_p\mathbf{1}_p^\top)) \leq \lambda_n, \\ &\hat{S}-\hat{L} \in \text{span}(\mathbf{1}_p\mathbf{1}_p^\top). \end{aligned} \tag{34}$$

Finally, Since $(S^\star - L^\star)\mathbf{1}_p\mathbf{1}_p^\top = 0$ and $\mathcal{A}(\Delta_S, \Delta_L)\mathbf{1}_p\mathbf{1}_p^\top = 0$, we have the following formulation of the optimality condition (34) in terms of the matrix $\mathbb{I}^\star$

$$\begin{aligned}
&\mathcal{P}_{\mathbb{H}}\mathcal{A}^\dagger(\mathbb{I}^\star(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top) + E_n) = Z,\\
&\Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger(\mathbb{I}^\star(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top) + E_n)) \leq \lambda_n, \qquad (35)\\
\hat{S} - \hat{L} \in \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top).&
\end{aligned}$$

It is straightforward to show that if for some $(\hat{S}, \hat{L})$, the second condition in (35) is satisfied with strict inequality, that is:

$$\Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger(\mathbb{I}^\star(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top) + E_n)) < \lambda_n.$$

H.4 Constrained optimization problem

We consider the following non-convex optimization problem:

$$\begin{aligned}
\underset{S\in\mathbb{S}^p, L\in\mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det(U^T(S-L)U) - \operatorname{tr}((S-L)\hat{\Gamma}_O/2) + \lambda_n(\|S\|_1 + \gamma\|L\|_\star),\\
\text{subject-to} \quad & S - L \in \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top)\ ;\ (S, L) \in \mathcal{M},
\end{aligned} \qquad (36)$$

where:

$$\mathcal{M} = \Bigg\{ S, L \in \mathbb{S}^p : S \in \Omega^\star, \operatorname{rank}(L) \leq \operatorname{rank}(L^\star)$$
$$\|\mathcal{P}_{T^{\star\perp}}(L - L^\star)\|_2 \leq \frac{C_0\lambda_n}{\psi} \ \ ; \ \ \Phi_\gamma(\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(S - S^\star, L - L^\star)) \leq C_0\lambda_n \Bigg\},$$

with $C_0 = 10 + \frac{32\sqrt{5h}}{\alpha'(1-\sqrt{1-(\kappa^{\star 2}-\omega)^2})(\frac{1}{\zeta}-2(\kappa^\star+\omega))}\left[1 + \frac{1}{3\zeta}\right]$. The optimization program (36) is non-convex due to the rank constraint $\operatorname{rank}(L) \leq \operatorname{rank}(L^\star)$ in the set $\mathcal{M}$. These constraints ensure that the matrix L belongs to an appropriate variety. The constraints in $\mathcal{M}$ along $T^{\star\perp}$ ensure that the tangent space $T(L)$ is close to $T^\star$. Finally, the last condition roughly controls the error. We begin by proving the following useful proposition:

Proposition 32 *Let (S, L) be a set of feasible variables of (36). Let $\Delta = (S - S^\star, L - L^\star)$. Then, $\Phi_\gamma(\Delta) \leq C_2\lambda_n$ where $C_2 = m\max\{\left(\frac{4C_0}{\alpha'} + \frac{1}{\psi}\right), 1\}$.*

Proof [Proof of Proposition 32] Let $\mathbb{H}^\star = \Omega^\star \times T^\star$. Then:

$$\begin{aligned}
\Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] &\leq \Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(\Delta)] + \Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)]\\
&\leq C_0\lambda_n + mC_0\lambda_n \leq 2mC_0\lambda_n.
\end{aligned}$$

Since $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\cdot)] \leq 2\Phi_\gamma[\cdot]$, we have that: $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq 4mC_0\lambda_n$. Then, appealing to Property $p1$ in (29), we have that: $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{4C_0\lambda_n}{\alpha'}$. Moreover, $\Phi_\gamma(\Delta) \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] +$

$\Phi_\gamma[\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] \leq \lambda_n m \left(\frac{4C_0}{\alpha'} + \frac{1}{\psi}\right)$. ■

Proposition 32 leads to powerful implications. In particular, under additional conditions on the minimum nonzero singular values of $L^\star$, any feasible set of variables (S, L) of (36) has two key properties: (a) The variables (S, L) are smooth points of their underlying varieties with $L \succeq 0$ and $S - L \succeq 0$, and (b) The constraints in $\mathcal{M}$ along $T^{\star\perp}$ are locally inactive at L. These properties, among others, are proved in the following corollary.

Corollary 33 *Consider any feasible variables (S, L) of (36). Let $T' = T(L)$. Let σ be the smallest nonzero singular value of $L^\star$ and s be the smallest in magnitude nonzero value of $S^\star$. Let $\mathbb{H}' = \Omega^\star \times T'$, $C_{T'} = \mathcal{P}_{T'^\perp}(L^\star)$ and $C_{T'\oplus \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top)} = \mathcal{P}_{(T'\oplus \mathrm{span}(\mathbf{1}_p\mathbf{1}_p^\top))^\perp}(L^\star)$. Suppose that the following inequalities are met: $\sigma \geq \max\left\{16m\bar{m}\frac{\lambda_n C_2}{\omega}, \frac{2\psi C_2^2\lambda_n}{C_0}, \left(mC_2 + \frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star+\omega))}{4\left[1+\frac{1}{3\zeta}\right]}\right)\lambda_n\right\}$ and $s \geq 12m\bar{m}\lambda_n C_2$. Then,*

1. *L and S are smooth points of their underlying varieties so that* $\mathrm{support}(\hat{S}) = \mathrm{support}(S^\star)$ *and* $\mathrm{rank}(\hat{L}) = \mathrm{rank}(L^\star)$. *Furthermore, $L \succeq 0$, and $S - L \succeq 0$*
2. $\|\mathcal{P}_{T^{\star\perp}}(\hat{L} - L^\star)\|_2 \leq \frac{C_0\lambda_n}{2\psi}$,
3. $\rho(T', T^\star) \leq \omega$,
4. $\max\{\Phi_\gamma(\mathcal{A}^\dagger\mathbb{I}^\star C_{T'}), \Phi_\gamma(\mathcal{A}^\dagger\mathbb{I}^\star C_{T'\oplus\mathbf{1}_p\mathbf{1}_p^\top})\} \leq \frac{\lambda_n}{6\zeta}$,
5. $\Phi_\gamma[\mathcal{A}^\dagger C_{T'}] \leq \frac{4\lambda_n}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star+\omega))}\left[1 + \frac{1}{3\zeta}\right]$.

Proof [Proof of Corollary 33] We appeal to the results regarding the perturbation analysis of the low-rank matrix variety.

1. Based on assumptions regarding the minimum nonzero singular value of $L^\star$ and minimum nonzero entry in magnitude of $S^\star$, one can check that since $\omega \leq 1$

$$\sigma \geq 12m\bar{m}\frac{\lambda_n C_2}{\omega} \geq 12m\bar{m}\lambda_n C_2 \geq 8\|L - L^\star\|_2,$$
$$s \geq 12m\bar{m}\lambda_n C_2 \geq 12m\bar{m}\lambda_n C_2 \geq 2\|S - S^\star\|_2.$$

   Combining these results, we conclude that S, L are smooth points of their varieties, namely that $\mathrm{rank}(L) = \mathrm{rank}(L^\star)$ and $\mathrm{support}(S) = \mathrm{support}(S^\star)$. The fact that $L \succeq 0$ follows from $\sigma \geq 2\|L - L^\star\|_2$. Furthermore, to check that $S - L \succeq 0$, first note that $\sigma_{\min}(S^\star - L^\star) \geq \frac{1}{\sqrt{\psi}}$. Then, $\|S - L - (S^\star - L^\star)\|_2 \leq 2mC_2\lambda_n$. From the choice of λ_n and the condition on the sample size, we have that $4mC_2\lambda_n < \frac{1}{\sqrt{\psi}}$. Thus, $S - L \succeq 0$.

2. Since $\sigma \geq 8\|L - L^\star\|_2$, we can appeal to Proposition 2.2 of Chandrasekaran et al. (2012) to conclude that the constrain5 in $\mathcal{M}$ along $\mathcal{P}_{T^{\star\perp}}$ is strictly feasible:

$$\|\mathcal{P}_{T^{\star\perp}}(L - L^\star)\|_2 \leq \frac{\|L - L^\star\|_2^2}{\sigma} \leq \frac{C_2^2\lambda_n^2}{\sigma} < \frac{C_0\lambda_n}{\psi}.$$

3. Appealing to Proposition 2.1 of Chandrasekaran et al. (2012), we prove that the tangent space T' is close to $T^\star$:

$$\rho(T', T^\star) \leq \frac{2\|L - L^\star\|_2}{\sigma} \leq \frac{2m\bar{m}\lambda_n C_2 \omega}{12 m\bar{m}\lambda_n C_2} \leq \omega.$$

4. Letting σ' be the minimum nonzero singular value of L. One can check that:

$$\sigma' \geq \sigma - \|L - L^\star\|_2 \geq \sigma - mC_2\lambda_n \geq 10mC_2\lambda_n \geq 8\|L - L^\star\|_2.$$

One can also obtain the following lower bounds for σ':

$$\sigma' \geq \sigma - \|L - L^\star\|_2 \geq \sigma - mC_2\lambda_n \geq 6\zeta m C_2^2 \psi \lambda_n - mC_2\lambda_n \geq 6\zeta m\psi C_2^2 \lambda_n$$
$$\sigma' \geq \sigma - \|L - L^\star\|_2 \geq \sigma - mC_2\lambda_n \geq \frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))\lambda_n}{4\left[1 + \frac{1}{3\zeta}\right]}$$

where we have used $C_2\psi \geq 1$. Once again appealing to Proposition 2.2 of Chandrasekaran et al. (2012) and simple algebra, we have:

$$\Phi_\gamma[\mathcal{A}^\dagger \mathbb{I}^\star C_{T'}] \leq m\psi \|C_{T'}\|_2 \leq m\psi \frac{\|L - L^\star\|_2^2}{\sigma'} \leq m\psi \frac{C_2^2\lambda_n^2}{6\zeta m\psi C_2^2 \lambda_n} \leq \frac{\lambda_n}{6\zeta}.$$

From Lemma 20, we have that $\|C_{T' \oplus \mathbf{1}_p\mathbf{1}_p^\top}\|_2 \leq \|C_{T'}\|_2$. Following the same logic as above, we can then show that: $\Phi_\gamma[\mathcal{A}^\dagger \mathbb{I}^\star C_{T' \oplus \mathbf{1}_p\mathbf{1}_p^\top}] \leq \frac{\lambda_n}{6\zeta}$.

5. Finally, we show that:

$$\Phi_\gamma[C_{T'}] \leq m\|\mathcal{P}_{T'^\perp}(L - L^\star)\|_2 \leq m\frac{\|L - L^\star\|_2^2}{\sigma'} \leq \frac{mC_2^2\lambda_n^2}{\sigma'} \leq \frac{4\lambda_n}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right].$$

■

Consider any optimal solution $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ of (36). We will show that $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ is the unique solution of the nonconvex program (36), as well as the unique solution of (30).

H.5 Variety constrained program to tangent space constrained program

Let $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ be any optimal solution of (36). In Corollary 33, we conclude that the variables $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ are smooth points of their respective varieties. As a result, the rank constraint $\text{rank}(L) \leq \text{rank}(L^\star)$ can be linearized to $L \in T(\hat{L}^{\mathcal{M}})$. Since all the remaining constraints are convex, the optimum of the linearized program is also the optimum of (36). Moreover, we once more appeal to Corollary 33 to conclude that the constraints in $\mathcal{M}$ along $T^{\star\perp}$ are strictly feasible at $\hat{L}^{\mathcal{M}}$. As a result, these constraints are inactive and can be removed in this "linearized program". We now argue that the constraint $\Phi_\gamma[\mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A}(\hat{S}^{\mathcal{M}} - S^\star, \hat{L}^{\mathcal{M}} - L^\star)]$ is

inactive. For notational simplicity, we let $T' = T(\hat{L}^{\mathcal{M}})$ and $\mathbb{H}' = \Omega^\star \times T'$, we consider the following optimization problem:

$$\begin{aligned}(\tilde{S}, \tilde{L}) = \underset{S\in\mathbb{S}^p, L\in\mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det(U^T(S-L)U) - \operatorname{tr}((S-L)\hat{\Gamma}_O/2) + \lambda_n(\|S\|_1 + \gamma\|L\|_\star), \\ \text{subject-to} \quad & (S,L) \in \mathbb{H}', S - L \in \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top).\end{aligned} \tag{37}$$

We prove that under conditions imposed on the regularization parameter λ_n, the pair of variables $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ is the unique optimum of (37). First, note that the optimum of (37) is unique since it is a strictly convex program convex because the negative log-likelihood terms have a strictly positive-definite Hessian due to property $p1$) in (29). To show that $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ is the optimum of (37), it suffices to show strict feasibility of the constraint, that is: $\Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}(\tilde{S}-S^\star, \tilde{L}-L^\star)] < C_0\lambda_n$.

From optimality conditions of (37), there exists $Q_\Omega \in \Omega^{\star\perp}$, $Q_T \in T'^{\perp}$, $t \in \mathbb{R}$ such that:

$$\begin{aligned} -\hat{\Gamma}_O/2 - (\tilde{S}-\tilde{L})^+ + t\mathbf{1}_p\mathbf{1}_p^\top + Q_\Omega &= -\lambda\partial\|\tilde{S}\|_1, \\ \hat{\Gamma}_O/2 + (\tilde{S}-\tilde{L})^+ - t\mathbf{1}_p\mathbf{1}_p^\top + Q_T &= -\lambda\gamma\partial\|\tilde{L}\|_\star, \\ \tilde{S}-\tilde{L} &\in \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top). \end{aligned} \tag{38}$$

Let the reduced SVD of $\tilde{L}$ be given by $\tilde{L} = \bar{U}\bar{D}\bar{V}^T$ and $Z = (\lambda\operatorname{sign}(\tilde{S}), \lambda\gamma\bar{U}\bar{V}^T)$. Following a similar logic as in Section H.3 and restricting the optimality conditions to the space of $\mathbb{H}$, we have the following equivalent characterization of the optimality conditions:

$$\begin{aligned}\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger(\mathbb{I}^\star(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top) + E_n) &= Z, \\ \tilde{S}-\tilde{L} &\in \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top).\end{aligned} \tag{39}$$

Here, $\Delta_S = \tilde{S} - S^\star$, $\Delta_L = \tilde{L} - L^\star$. In the remaining, we will denote $\Delta_{L+} = \tilde{L} - L^\star + t\mathbf{1}_p\mathbf{1}_p^\top$. Our result relies on the following propositions to control the remainder term.

Proposition 34 *Suppose* $\Phi_\gamma(\Delta_S, \Delta_{L+}) \leq \frac{1}{2C_1}$ *for* $C_1 = \psi(m + d^\star)$ *and any* $\Delta_S \in \Omega^\star$. *Then,* $\Phi_\gamma[\mathcal{A}^\dagger\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\Delta_S, \Delta_{L+}))] \leq 2m\psi C_1^2\Phi_\gamma(\Delta_S, \Delta_{L+})^2$.

Proof [Proof of Proposition 34] We have that:

$$\begin{aligned}\|\mathcal{A}(\Delta_S, \Delta_{L+})\|_2 \leq \|\Delta_S\|_2 + \|\Delta_{L+}\|_2 &\leq \theta(\Omega^\star)\|\Delta_S\|_\infty + \gamma\frac{\|\Delta_{L+}\|_2}{\gamma} \leq (\gamma + \theta(\Omega^\star))\Phi_\gamma(\Delta_S, \Delta_{L+}) \\ &\leq (m + d^\star)\Phi_\gamma(\Delta_S, \Delta_{L+}) \leq \frac{1}{2\psi}.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\Delta_S, \Delta_{L+}))\|_2 &\leq \psi\sum_{k=2}^{\infty}(\|\Delta_S + \Delta_{L+}\|_2\psi)^k \leq \psi^3\|\Delta_S + \Delta_{L+}\|_2^2\frac{1}{1 - \|\Delta_S + \Delta_{L+}\|_2\psi} \\ &\leq 2\psi^3\left(1 + \frac{\alpha'}{6\zeta}\right)^2\Phi_\gamma(\Delta_S, \Delta_{L+})^2 = 2\psi C_2^2\Phi_\gamma(\Delta_S, \Delta_{L+})^2.\end{aligned}$$

Putting everything together, we have the desired result. ■

Notice that the bound on the remainder term is dependent on the error term $\Phi_\gamma(\Delta_S, \Delta_{L+})$. In the following proposition, we bound this error so we can control the remainder term.

Proposition 35 *Let $\tilde{S}, \tilde{L}$ be the solution of convex program* (37)*. Define*

$$r = \max\left\{ \frac{4}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))} [\Phi_\gamma(\mathcal{A}^\dagger E_n) + \Phi_\gamma(\mathcal{A}^\dagger \mathbb{I}^\star C_{T'}) + \lambda_n], \Phi_\gamma[(0, C_{T'})] \right\}.$$

If we have that $r \leq \min\left\{ \frac{8\alpha'}{C_1}, \frac{\min\{\alpha', 1\}(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{16 m \psi C_2^2} \right\}$, then $\Phi_\gamma(\Delta_S, \Delta_L) \leq \frac{4r\sqrt{5h}}{1 - \sqrt{1 - (\kappa^{\star 2} - \omega)^2}}$ and $\Phi_\gamma(0, t\mathbf{1}_p\mathbf{1}_p^\top) \leq \frac{4r\sqrt{5h}}{1 - \sqrt{1 - (\kappa^{\star 2} - \omega)^2}}$.

The proof of the proposition relies on the following lemma which we state and prove first.

Lemma 36 *Consider the following optimization:*

$$\begin{aligned} &\underset{S \in \mathbb{S}^p, L \in \mathbb{S}^p}{\operatorname{argmin}} \ \log\det(U^T(S-L)U) - \operatorname{tr}((S-L)\hat{\Gamma}_O/2) + \operatorname{tr}(\mathbf{1}_p\mathbf{1}_p^\top(S-L)) + \lambda_n(\|S\|_1 + \gamma\|L\|_\star). \\ &\textit{subject-to} \quad (S, L) \in \mathbb{H}' \end{aligned} \tag{40}$$

Then, the solution of (40) *is unique and is equal to $\tilde{S}, \tilde{L}$ (i.e. the solution of* (37)*).*

Proof [Proof of Lemma 36] Note that by property $p1$) in (29), the estimator (40) is strictly convex. We will denote the optimal solution of (40) by $(\tilde{S}, \tilde{L})$. We are using the same notation as the optimal solution of (37) as we will show momentarily that these optimal solutions are identical. Specifically, define Z as is done before Proposition 34. Let $\Delta_S = \tilde{S} - S^\star$ and $\Delta_L = \tilde{L} - L^\star$. The optimality condition of (40) is given by:

$$\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger(\mathbb{I}^\star\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top) + E_n) = Z. \tag{41}$$

Notice that the optimality condition (41) is identical to the first condition in (39). Since (40) has a unique solution, then, the optimal solutions of (37) and (40) coincide. ■

Proof [Proof of Proposition 35] Since T' is a tangent space such that $\rho(T', T^\star) \leq \omega$, we have from Property $p1$ in (29) that the operator $\mathcal{B} = (\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger\mathbb{I}^\star\mathcal{A}\mathcal{P}_{\mathbb{H}'})^{-1}$ is bijective and is well-defined. Consider the following function taking as input $(\delta_S, \delta_{L+}) \in \mathbb{Q}'$ where $\mathbb{Q}' = \Omega^\star \times (T' \oplus t\mathbf{1}_p\mathbf{1}_p^\top)$:

$$F(\delta_S, \delta_{L+}) = (\delta_S, \delta_{L+}) - \mathcal{B}\left\{ \mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger[\mathbb{I}^\star\mathcal{A}(\delta_S, \delta_{L+}) + \mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\delta_S, \delta_{L+} + C_{T'})) + \mathbb{I}^\star C_{T'} + E_n - Z \right\}.$$

Here, $C_{T'} = \mathcal{P}_{T'^\perp}(L^\star)$. Now a point (δ_S, δ_{L+}) is a fixed point of F if and only if $\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger[\mathbb{I}^\star\mathcal{A}(\delta_S, \delta_{L+}) + \mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\delta_S, \delta_{L+} + C_{T'})) + \mathbb{I}^\star C_{T'} + E_n] = Z$. Further, a fixed point (δ_S, δ_{L+}) provides

certificates of optimality for (40). Specifically, let $\tilde{S} = S^\star + \delta_S$. By Lemma 21, find a unique decomposition of $\delta_{L+} = L + t\mathbf{1}_p\mathbf{1}_p^\top$ where $L \in T'$. Then, let $\tilde{L} = \mathcal{P}_{T'}(L^\star) + L$. By construction, the parameters $(\tilde{S}, \tilde{L})$ then satisfy the optimality condition for (41) and thus also the optimality condition of (39) after appealing to Lemma 36. In other words, the fixed point of the function F is $\mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L) + (0, t\mathbf{1}_p\mathbf{1}_p^\top)$.

Next, using Brouwer's fixed point theorem, we show that F has a fixed point that lies in the ball $\mathbb{B}_r = \{(\delta_S, \delta_{L+}) \in \mathbb{Q}' | \Phi_\gamma(\delta_S, \delta_{L+}) \leq r\}$. An equivalent formulation of F is:

$$F(\delta_S, \delta_{L+}) = \mathcal{P}_{\mathbb{H}'^\perp}(\delta_S, \delta_{L+}) - \mathcal{B}\Big\{\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger[\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\delta_S, \delta_{L+} + C_{T'})) + \mathbb{I}^\star[C_{T'} + \mathcal{A}\mathcal{P}_{\mathbb{H}'^\perp}(\delta_S, \delta_{L+})] + E_n - Z\Big\}.$$

First, note that by appealing to Lemma 19, we have that: $\Phi_\gamma\left[\mathcal{P}_{\mathbb{H}'^\perp}(\delta_S, \delta_{L+})\right] \leq 2r(\kappa^\star + \omega)$. Similarly, we have from Property $p3$ in (29) that: $\Phi_\gamma\left[\mathcal{B}\left\{\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger I\mathcal{A}\mathcal{P}_{\mathbb{H}'^\perp}(\delta_S, \delta_{L+})\right\}\right] \leq r\left(1 - \frac{1}{\zeta}\right)$. Finally, we note that:

$$\begin{aligned}
&\Phi_\gamma\left[\mathcal{B}\left\{\mathcal{P}_{\mathbb{H}'}\mathcal{A}^\dagger[\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\delta_S, \delta_{L+} + C_{T'})) + \mathbb{I}^\star C'_T + E_n - Z\right\}\right] \\
&\leq \frac{2}{\alpha'}\left(\Phi_\gamma[\mathcal{A}^\dagger\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\delta_S, \delta_{L+} + C_{T'}))] + \Phi_\gamma[\mathbb{I}^\star C_{T'}] + \Phi_\gamma[E_n] + \lambda_n\right) \\
&\leq \frac{r(\frac{1}{\zeta} - 2(\kappa^\star - \omega))}{2} + \frac{2}{\alpha'}\left(\Phi_\gamma[\mathcal{A}^\dagger\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}(\delta_S, \delta_{L+} + C_{T'}))]\right)
\end{aligned}$$

where the last inequality is by the definition of r. By the assumption on r, we have that $\Phi_\gamma((\delta_S, \delta_{L+}) + (0, C_{T'})) \leq \frac{1}{2C_1}$. And so we can appeal to Proposition 34 to conclude that:

$$\frac{2}{\alpha'}\Phi_\gamma[\mathcal{A}^\dagger\mathcal{R}_{\Gamma_0^\star}(\mathcal{A}((\delta_S, \delta_{L+} + C_{T'}) \leq \frac{8m\psi C_1^2 r^2}{\alpha'} \leq \frac{16m\psi C_2^2 r}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\frac{r(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{2} \leq r/2,$$

where the last inequality uses the bound on r. So by Brouwer's fixed point theorem, we conclude that: $\Phi_\gamma[\mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L) + (0, t\mathbf{1}_p\mathbf{1}_p^\top)] \leq r$. Finally, note that: $\Phi_\gamma[\mathcal{P}_{\mathbb{H}'^\perp}(\Delta_S, \Delta_L)] \leq r$. Thus, $\Phi_\gamma[(\Delta_S, \Delta_L) + (0, t\mathbf{1}_p\mathbf{1}_p^\top)] \leq 2r$. Finally, appealing to Lemma 22 and some manipulations, we have the bound $\max\{\Phi_\gamma(\Delta_S, \Delta_L), t\mathbf{1}_p\mathbf{1}_p^\top\} \leq \frac{4r\sqrt{5h}}{1-\sqrt{1-(\kappa^{\star 2}-\omega)^2}}$. ■

Proposition 37 *Suppose that* $\Phi_\gamma[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\zeta}$ *and suppose that:*

$$\lambda_n \leq \min\left\{\min\left\{\frac{8\alpha'}{C_1}, \frac{\min\{\alpha', 1\}(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{16m\psi C_2^2}\right\}\frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{4(1 + \frac{1}{3\zeta})}, \frac{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{64C_1(1 + \frac{1}{3\zeta})}, \frac{\alpha'^2(\frac{1}{\zeta} - 2(\kappa^\star + \omega))^2}{6144\zeta(1 + \frac{1}{3\zeta})^2}\right\}.$$

Then, we have that: $\tilde{S} = \hat{S}^{\mathcal{M}}$, $\tilde{L} = \hat{L}^{\mathcal{M}}$.

Proof From Corollary 33, we have that $\Phi_\gamma[\mathcal{A}^\dagger \mathbb{I}^\star C_{T'}] \leq \frac{\lambda_n}{6\zeta}$. We then have that:

$$\begin{aligned}
&\frac{4}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}[\Phi_\gamma(\mathcal{A}^\dagger E_n) + \Phi_\gamma(\mathcal{A}^\dagger \mathbb{I}^\star C_{T'}) + \lambda_n] \\
&\leq \frac{4\lambda_n}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right] \leq \min\left\{\frac{8\alpha'}{C_1}, \frac{\min\{\alpha', 1\}(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}{16 m \psi C_2^2}\right\}.
\end{aligned}$$

We also have from Corollary 33 that $\Phi_\gamma(\mathcal{A}^\dagger C_{T'}) \leq \frac{4\lambda_n}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right]$. Let $r = \frac{4\lambda_n}{\alpha'(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right]$. We can appeal to Proposition 35 to conclude that:

$$\Phi_\gamma[\Delta_S, \Delta_L] \leq \frac{16\lambda_n\sqrt{5h}}{\alpha'(1 - \sqrt{1 - (\kappa^{\star 2} - \omega)^2})(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right].$$

From the bound on λ_n, we have that: $\Phi_\gamma[\Delta_S, \Delta_L] \leq \frac{1}{2C_1}$. So we can appeal to Proposition 34 to conclude that:

$$\Phi_\gamma[\mathcal{A}^\dagger \mathcal{R}_{\Gamma_O^\star} \mathcal{A}(\Delta_S, \Delta_L)] \leq 2m\psi C_1^2 \Phi_\gamma[\Delta_S, \Delta_L]^2 \leq \frac{\lambda_n}{6\zeta}, \tag{42}$$

where here again we use the bound on λ_n. Note that $\Delta_{L+} = \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top$. We have from Corollary 33 that $\Phi_\gamma[\mathcal{A}^\dagger \mathbb{I}^\star C_{T'}] \leq \frac{\lambda_n}{6\zeta}$. From the optimality conditions of (37), we have that:

$$\begin{aligned}
&\Phi_\gamma(\mathcal{P}_{\mathbb{H}'} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L)) \\
&\leq 2\lambda_n + 2\Phi_\gamma(0, t\mathbf{1}_p\mathbf{1}_p^\top) + \Phi_\gamma[\mathcal{A}^\dagger \mathcal{R}_{\Gamma_O^\star} \mathcal{A}(\Delta_S, \Delta_L)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}'} \mathcal{A}^\dagger \mathbb{I}^\star C_{T'}] + \Phi_\gamma[\mathcal{A}^\dagger E_n], \\
&\leq 2\lambda_n + \frac{\lambda_n}{2\zeta} + \frac{16\lambda_n\sqrt{5h}}{\alpha'(1 - \sqrt{1 - (\kappa^{\star 2} - \omega)^2})(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right],
\end{aligned}$$

where the second inequality follows from bound on $\Phi_\gamma((0, t\mathbf{1}_p\mathbf{1}_p^\top))$ in Proposition 35. Appealing to property $p2$ in (29): $\Phi_\gamma(\mathcal{P}_{\mathbb{H}'^\perp} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L)) \leq \Phi_\gamma(\mathcal{P}_{\mathbb{H}'^\perp} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L))$. Thus

$$\begin{aligned}
\Phi_\gamma(\mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A}(\Delta_S, \Delta_L)) &\leq \Phi_\gamma(\mathcal{P}_{\mathbb{H}'} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L)) + \Phi_\gamma(\mathcal{P}_{\mathbb{H}'^\perp} \mathcal{A}^\dagger \mathbb{I}^\star \mathcal{A} \mathcal{P}_{\mathbb{H}'}(\Delta_S, \Delta_L)) \\
&\quad + \Phi_\gamma[\mathcal{A}^\dagger \mathbb{I}^\star C_{T'}] \leq 8\lambda_n + \frac{32\lambda_n\sqrt{5h}}{\alpha'(1 - \sqrt{1 - (\kappa^{\star 2} - \omega)^2})(\frac{1}{\zeta} - 2(\kappa^\star + \omega))}\left[1 + \frac{1}{3\zeta}\right] \\
&\quad < C_0\lambda_n.
\end{aligned}$$

■

H.6 Removing the tangent space constraint

It remains to connect the estimator (37) with (9). In particular, we check that $\tilde{S} = \hat{S}$ and $\tilde{L} = \hat{L}$ where $(\tilde{S}, \tilde{L})$ is the solution of (37) and $(\hat{S}, \hat{L})$ is the solution of (9). We formalize this in the following proposition.

Proposition 38 *Suppose that* $\Phi_\gamma[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\zeta}$. *Then,* $\tilde{S} = \hat{S}$ *and* $\tilde{L} = \hat{L}$.

Proof [Proof of Proposition 38] We must show that $(\tilde{S}, \tilde{L})$ satisfy the optimality conditions of (30) in (35), namely that there exists a dual variable t such that

$$\begin{aligned} \mathcal{P}_{\mathbb{H}}\mathcal{A}^\dagger(\mathbb{I}^\star(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L) + E_n) = Z, \\ \Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger(\mathbb{I}^\star(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) + \mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L) + E_n)) < 1, \\ \tilde{S} - \tilde{L} \in \operatorname{span}(\mathbf{1}_p\mathbf{1}_p^\top), \end{aligned} \tag{43}$$

where $\Delta_S = \tilde{S} - S^\star$ and $\Delta_L = \tilde{L} - L^\star$. Notice that the first and third optimality conditions are the same as (39). It remains to show the second inequality where the strict inequality is to ensure that $(\tilde{S}, \tilde{L})$ is the unique solution. It suffices to show that:

$$\begin{aligned} &\Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger(\mathbb{I}^\star\mathcal{P}_{\mathbb{Q}'}(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) \\ &< \lambda_n - \Phi_\gamma[\mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L)] - \Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star C_{T' \oplus \mathbf{1}_p\mathbf{1}_p^\top}] - \Phi_\gamma[\mathcal{A}^\dagger E_n]. \end{aligned} \tag{44}$$

Manipulating the first optimality condition, we have that:

$$\begin{aligned} \Phi_\gamma(\mathcal{P}_{\mathbb{H}}\mathcal{A}^\dagger(\mathbb{I}^\star\mathcal{P}_{\mathbb{Q}'}(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) \leq \lambda_n + 2(\Phi_\gamma[\mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L)] + \Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star C_{T' \oplus \mathbf{1}_p\mathbf{1}_p^\top}] \\ + \Phi_\gamma[\mathcal{A}^\dagger E_n]) \leq \lambda_n + \frac{\lambda_n}{\zeta} = \lambda_n\left(1 + \frac{1}{\zeta}\right), \end{aligned}$$

where we have here used the bound $\Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star C_{T' \oplus \mathbf{1}_p\mathbf{1}_p^\top}] \leq \frac{\lambda_n}{6\zeta}$ from Corollary 33 and the bounds $\Phi_\gamma[\mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L)] \leq \frac{\lambda_n}{6\zeta}$ from (42) and $\Phi_\gamma[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\zeta}$ from proposition statement. Appealing to property $p2$ in (29), we then have that:

$$\begin{aligned} \Phi_\gamma(\mathcal{P}_{\mathbb{H}^\perp}\mathcal{A}^\dagger(\mathbb{I}^\star\mathcal{P}_{\mathbb{Q}'}(\mathcal{A}(\Delta_S, \Delta_L + t\mathbf{1}_p\mathbf{1}_p^\top)) &\leq \lambda_n\left(1 + \frac{1}{\zeta}\right)\left(1 - \frac{1}{\zeta}\right) \\ &= \lambda_n\left(1 - \frac{1}{\zeta^2}\right) < \lambda_n\left(1 - \frac{1}{2\zeta}\right). \end{aligned}$$

Since $\Phi_\gamma[\mathcal{R}_{\Gamma_0^\star}\mathcal{A}(\Delta_S, \Delta_L)] + \Phi_\gamma[\mathcal{A}^\dagger\mathbb{I}^\star C_{T' \oplus \mathbf{1}_p\mathbf{1}_p^\top} + \Phi_\gamma[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{2\zeta}$, (44) holds.

■

H.7 Bounding the error term $\Phi_\gamma[\mathcal{A}^\dagger E_n]$

Let $\lambda_n = C_5 \left[\frac{24m\zeta}{\sqrt{c_5}} \sqrt{\frac{p^2 \log(\tilde{C}_5 p)}{k}} + \frac{6m\zeta}{\sqrt{k}} \right]$ where $c_5, C_5, \tilde{C}_5$ are defined in Theorem 30.

Lemma 39 *Under the conditions of Theorem 9, we have:*

$$\mathbb{P}\left(\Phi_\gamma[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\zeta} \right) \geq 1 - p^{-1}.$$

Proof Note that $\Phi_\gamma[\mathcal{A}^\dagger E_n] \leq m\|\Gamma^\star_O - \hat{\Gamma}_O\|_2 \leq pm\|\Gamma^\star_O - \hat{\Gamma}_O\|_\infty$. To show that, $\Phi_\gamma[\mathcal{A}^\dagger E_n] \leq \frac{\lambda_n}{6\zeta}$, it suffices to show that

$$\|\Gamma^\star_O - \hat{\Gamma}_O\|_\infty \leq \frac{4C_5}{\sqrt{c_5}} \sqrt{\frac{\log(\tilde{C}_5 p)}{k}} + \frac{C_5}{\sqrt{k}}. \tag{45}$$

Based on the condition on k, it is straightforward to show that:

$$C_5 \left\{ \left(\frac{k}{n}\right)^\xi (\log(n/k))^2 + \frac{1+\vartheta}{\sqrt{k}} \right\} \leq \frac{4C_5}{\sqrt{c_5}} \sqrt{\frac{\log(\tilde{C}_5 p)}{k}} + \frac{C_5}{\sqrt{k}}.$$

for $\vartheta = 2\sqrt{\log(\tilde{C}_5 p)}/\sqrt{c_5}$. Note that $\vartheta \leq \sqrt{k}/\log(n)^4$. Furthermore, $k \leq n/2$. Appealing to Corollary 30, we have that with probability greater than $1 - \tilde{C}_5 p^3 e^{-c_5 \vartheta^2} = 1 - p^{-1}$ that the bound in (45) is satisfied. ■

H.8 Summary and putting things together

Combining Propositions 37-38, we conclude that under the conditions of Theorem 9, with probability greater than $1 - 1/p$, the optimal solution $(\hat{S}, \hat{L})$ of (30) is unique and equal to an optimal solution $(\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ of (36). From Corollary 33, we have that $\hat{S} - \hat{L} \succeq 0$, $\hat{L} \succeq 0$. Thus, $(\hat{S}, \hat{L}) = (\hat{S}^{\mathcal{M}}, \hat{L}^{\mathcal{M}})$ is also the unique minimizer of (9). The guarantees on the closeness of $(\hat{S}, \hat{L})$ to the population parameters $(S^\star, L^\star)$ follow from Corollary 33 and Proposition 32.

Appendix I. Refitting for `eglatent`

Suppose $(\hat{S}, \hat{L})$ is the solution of (9) in the first step. We then obtain refitted parameters $(\tilde{S}, \tilde{L})$ as the second step by solving the following convex optimization program:

$$\begin{aligned} (\tilde{S}, \tilde{L}) = \underset{S \in \mathbb{S}^p, L \in \mathbb{S}^p}{\operatorname{argmin}} \quad & -\log\det(U^T(S-L)U) - \operatorname{tr}((S-L)\hat{\Gamma}_O/2), \\ \text{s.t.} \quad & S - L \succeq 0, L \succeq 0, (S-L)\mathbf{1}_p = 0, \\ & \operatorname{support}(S) \subseteq \operatorname{support}(\hat{S}), \text{col-space}(L) \subseteq \text{col-space}(\hat{L}). \end{aligned}$$

Here, the constraint support$(S) \subseteq$ support$(\hat{S})$ restricts the graph structure of our refitted solution to be contained in the graph estimated in the first step. Similarly, the constraint col-space$(L) \subseteq$ col-space$(\hat{L})$ restricts the row/column space of the refitted low-rank term to be contained in the row/column space estimated in the first step.

Appendix J. Additional experimental results

J.1 Synthetic experiments on different graph structure

We consider the exact same setup as in the simulation study in Section 5.1.1. The only difference is that we specify the sub-graph $\mathcal{G}_0 = (E_O, O)$ among the observed variables to be an Erdős–Rënyi with edge probability 0.08 and set $\Theta^\star_{ij}$ to -2 for every $(i, j) \in E_O$ and zero otherwise. The rest of the simulation study is carried out as described in Section 5.1.1. Figure 7 summarizes the performance of all the methods on 50 independent results. We again observe that our approach outperforms `eglearn`, and accurately recovers the graphical structure among the observed variables as well as the number of latent variables. In terms of validation likelihood, `eglatent` is a bit weaker than in the simulation with the cycle graph.

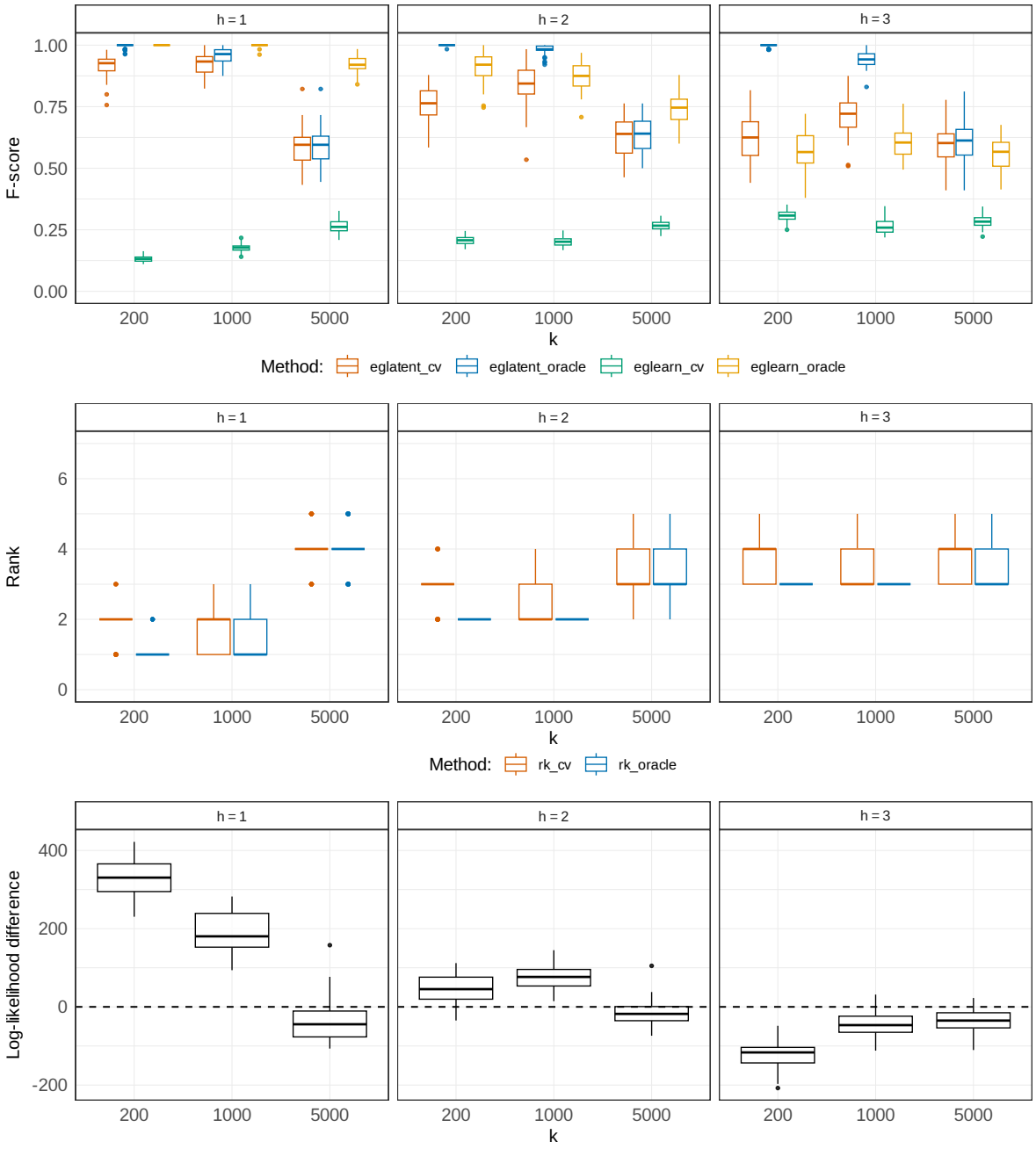

Figure 7: F-score (top row) and estimated number of latent variables (middle row) of `eglatent` method with the selection of the tuning parameter based on the oracle and validation on the F-score for the random graph with $h = 1, 2, 3$ latent variables and different effective sample sizes $k = 200, 1000, 5000$. The bottom row shows the difference between best `eglatent` and best `eglearn` log-likelihoods on the validation set.

J.2 Synthetic experiments on different values of γ

We consider the exact same setup as in the simulation study in Section 5.1.1. The only difference is the values of γ that are used in the `eglatent` estimator. We generate $k = 1000$ effective samples. Figure 8 shows the performance of `eglatent` for $\gamma \in \{2, 4, 6\}$. We observe that the performance of `eglatent` does not vary drastically with changes in γ, and continues to perform better than `eglearn`, especially for $h \in \{1, 2\}$. We also notice that $\gamma = 4$ yields the best-validated model for $h \in \{1, 2, 3\}$, hence why this value was chosen in our experiments in Section 5.1.1.

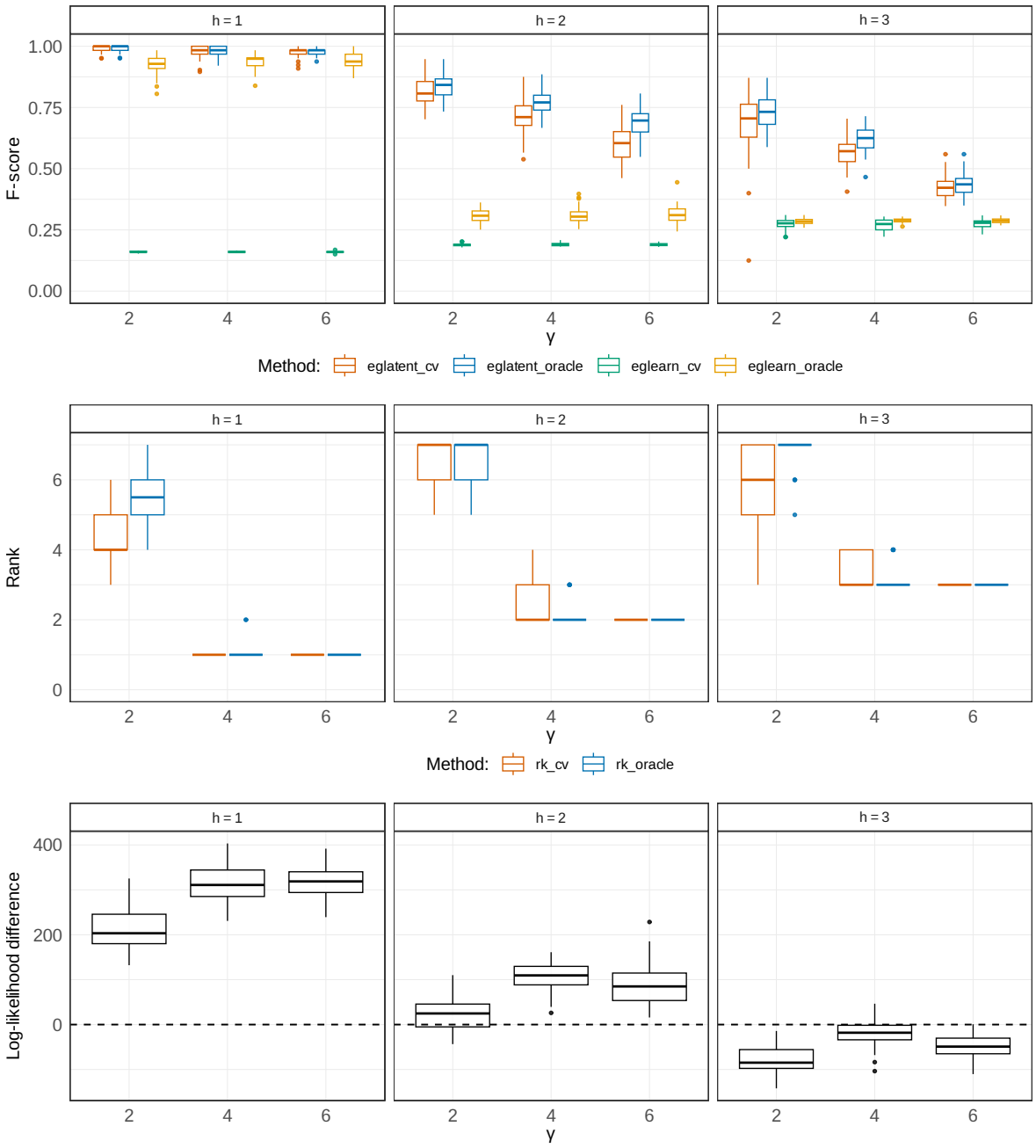

Figure 8: F-score (top row) and estimated number of latent variables (middle row) of `eglatent` method with the selection of the tuning parameter based on the oracle and validation on the F-score for the cycle graph with $h = 1, 2, 3$ latent variables and different regularization parameter $\gamma = 2, 4, 6$. The bottom row shows the difference between best `eglatent` and best `eglearn` log-likelihoods on the validation set. The effective sample size is set to $k = 1000$.

J.3 Synthetic experiments on comparison to the performance of Gaussian latent variable graphical model estimator

We compare the performance of our `eglatent` estimator to the Gaussian latent variable graphical model estimator in Chandrasekaran et al. (2012) (denoted by `LVGM`). We generate the data according to the setting in Appendix J.1. As the approach in Chandrasekaran et al. (2012) assumes Gaussian data, we transform the marginal distributions of each variable to standard normal distribution, before supplying the data to the Gaussian estimator. The following table compares the performance of the two estimators, where 'CV' is when the regularization parameters are chosen via the validation set, and 'Oracle' is when the regularization parameters are chosen to obtain the best F-score.

Table 2: Perfomance of `eglatent` compared with Gaussian estimator in Chandrasekaran et al. (2012)

	Oracle `eglatent`		CV `eglatent`		Oracle `LVGM`		CV `LVGM`	
# latents (h)	F-score	$\hat{h}$	F-score	$\hat{h}$	F-score	$\hat{h}$	F-score	$\hat{h}$
$h=1$	0.94(±0.02)	1.58(±0.53)	0.92(±0.04)	1.68(±0.55)	0.08(±0.04)	1.68(±1.88)	0.06(±0.03)	8.1(±0.83)
$h=2$	0.97(±0.01)	2(±0)	0.84(±0.07)	2.48(±0.54)	0.07(±0.04)	4.94(±3.01)	0.05(±0.04)	7.94(±0.86)
$h=3$	0.93(±0.03)	3(±0)	0.70(±0.08)	3.42(±0.57)	0.06(±0.03)	4.58(±2.39)	0.05(±0.03)	8.41(±0.94)

J.4 Additional results concerning the application

We report here the results of the application in Section 5.2. For thresholds $q = 0.85$ and $q = 0.95$, Figures 9 and 10 show the number of edges of `eglatent` and of `eglearn` and the validation log-likelihood values as a function of the tuning parameter λ_n. Figure 11 compares the different estimated graphs among the observed variables for the three thresholds.

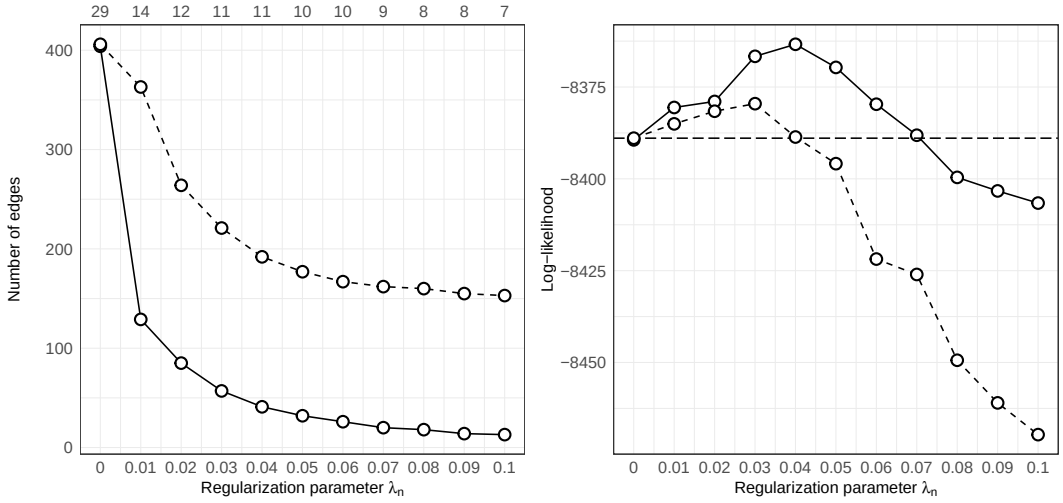

Figure 9: Results for threshold $q = 0.85$. Left: number of edges of the estimated graph of `eglearn` (dashed line) and the estimated sub-graph of observed variables of `eglatent` (solid line) as functions of the regularization parameter ρ; top axis shows the number of latent variables in `eglatent`. Right: corresponding log-likelihoods; horizontal line is the validation log-likelihood of the fully connected graph.

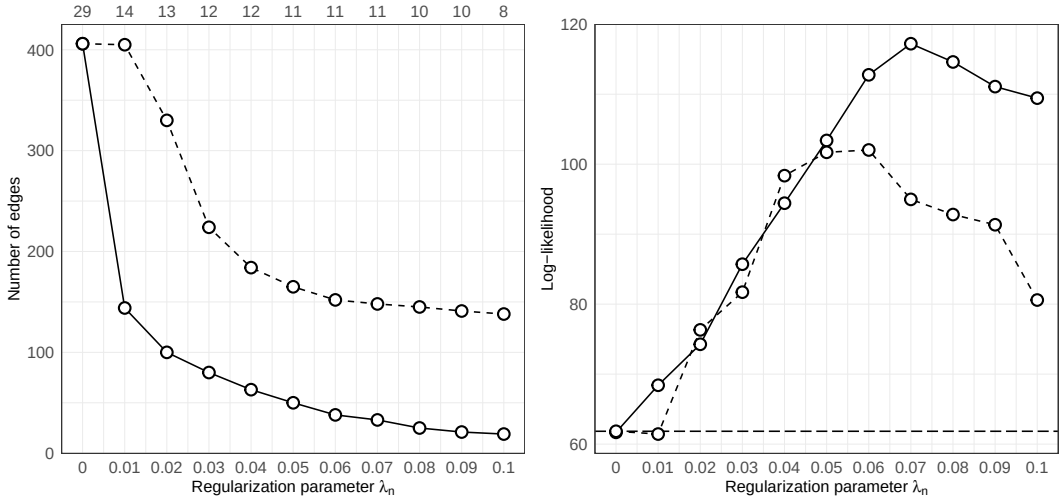

Figure 10: Results for threshold $q = 0.95$. Left: number of edges of the estimated graph of `eglearn` (dashed line) and the estimated sub-graph of observed variables of `eglatent` (solid line) as functions of the regularization parameter ρ; top axis shows the number of latent variables in `eglatent`. Right: corresponding log-likelihoods; horizontal line is the validation log-likelihood of the fully connected graph.

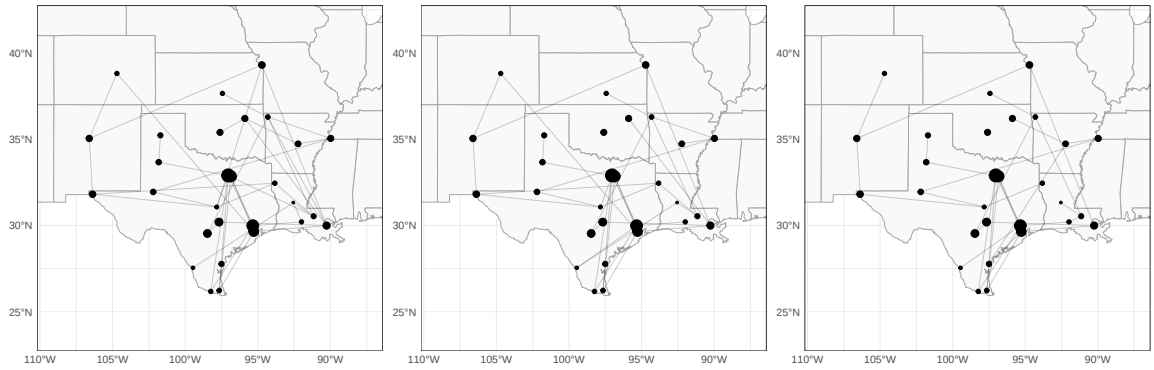

Figure 11: Airports in the Southern U.S. (dots) and flight connections, where the thickness of the nodes indicates the average number of daily flights at the airports. Estimated sub-graphs corresponding to observed variables of optimal `eglatent` models for exceendance thresholds 0.85 (left), 0.90 (center) and 0.95 (right).

References

R. Albert and A. Barabási. Statistical mechanics of complex networks. cond-mat/0106096, 2001. URL https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.74.47.

P. Asadi, A. C. Davison, and S. Engelke. Extremes on river networks. *Annals of Applied Statistics*, 9:2023–2050, 2015. URL https://www.jstor.org/stable/43826454.

S. Asenova and J. Segers. Extremes of Markov random fields on block graphs: max-stable limits and structured Hüsler–Reiss distributions. *Extremes*, 26:433–468, 2023. URL https://link.springer.com/article/10.1007/s10687-023-00467-9.

S. Asenova, G. Mazo, and J. Segers. Inference on extremal dependence in the domain of attraction of a structured Hüsler–Reiss distribution motivated by a Markov tree with latent variables. *Extremes*, 24:461–500, 2021. URL https://link.springer.com/article/10.1007/s10687-021-00407-5.

S.P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundational Trends of Machine Learning*, 3:1–122, 2011. URL https://dl.acm.org/doi/10.1561/2200000016.

E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 55(6):111–119, 2012. URL https://link.springer.com/article/10.1007/s10208-009-9045-5.

E Candès, X Li, Y Ma, and J Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011. URL https://dl.acm.org/doi/10.1145/1970392.1970395.

V. Chandrasekaran, V. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal of Optimization*, 21:572–596, 2011. URL https://epubs.siam.org/doi/10.1137/090761793.

V. Chandrasekaran, P. Parillo, and A. Willsky. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40:1935–1967, 2012. URL https://www.jstor.org/stable/41806519.

A. Chang and G.I. Allen. Subbotin graphical models for extreme value dependencies with applications to functional neuronal connectivity. *Annals of Applied Statistics*, 17(3):2364–2386, 2023. URL https://doi.org/10.1214/22-aoas1723.

L. de Haan and S.I. Resnick. Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 40:317–337, 1977. URL https://link.springer.com/article/10.1007/BF00533086.

C. Dombry, S. Engelke, and M. Oesting. Exact simulation of max-stable processes. *Biometrika*, 103:303–317, 2016. URL https://academic.oup.com/biomet/article/103/2/303/1744000.

S. Engelke and A.S. Hitz. Graphical models for extremes (with discussion). *Journal of the Royal Statistical Society Series B Stat. Methodol*, 82(4):871–932, 2020. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12355.

S. Engelke and J. Ivanovs. Sparse structures for multivariate extremes. *Annual Review of Statistics and Applications*, 8:241–270, 2021. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-040620-041554.

S. Engelke and S. Volgushev. Structure learning for extremal tree models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2055–2087, 2022. ISSN 1369-7412. doi: 10.1111/rssb.12556. URL https://doi.org/10.1111/rssb.12556.

S. Engelke, T. Opitz, and J. Wadsworth. Extremal dependence of random scale constructions. *Extremes*, 22:623–666, 2019. URL https://link.springer.com/article/10.1007/s10687-019-00353-3.

S. Engelke, A.S. Hitz, N. Gnecco, and M. Hentschel. graphicalExtremes: Statistical methodology for graphical extreme value models. *R Package Version 0.3.2*, 2022a.

S. Engelke, J. Ivanovs, and K. Strokorb. Graphical models for infinite measures with applications to extremes and Lévy processes, 2022b. URL https://arxiv.org/abs/2211.15769.

S. Engelke, M. Lalancette, and S. Volgushev. Learning extremal graphical structures in high dimensions, 2022c. URL https://arxiv.org/abs/2111.00840.

S. Engelke, M. Hentschel, Michaël Lalancette, and F. Röttger. Graphical models for multivariate extremes. 2024a. URL https://arxiv.org/abs/2402.02187.

S. Engelke, J. Ivanovs, and J.D. Thøstesen. Lévy graphical models, 2024b. URL https://arxiv.org/abs/2410.19952.

M. Fazel, H.A. Hindi, and S.P. Boyd. Rank minimization and applications in system theory. *Proceedings of the 2004 American Control Conference*, 4:3273–3278 vol.4, 2004. URL https://ieeexplore.ieee.org/document/1384521.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007. ISSN 1465-4644. URL https://doi.org/10.1093/biostatistics/kxm045.

M. Hentschel, S. Engelke, and J. Segers. Statistical inference for Hüsler–Reiss graphical models through matrix completions, 2022. URL https://arxiv.org/abs/2210.14292.

S. Hu, Z. Peng, and J. Segers. Modelling multivariate extreme value distributions via markov trees, 2022. URL https://arxiv.org/abs/2208.02627.

J. Hüsler and R. Reiss. Maxima of normal random vectors: Between independence and complete dependence. *Statist. Prob. Letters*, 7(4):283–286, February 1989. URL https://ideas.repec.org/a/eee/stapro/v7y1989i4p283-286.html.

S.L. Lauritzen. *Graphical models*, volume 17 of *Oxford statistical science series*. Clarendon Press, Oxford, 1996. ISBN 0198522193. URL https://www.tib.eu/de/suchen/id/TIBKAT%3A197598226.

J. Lederer and M. Oesting. Extremes in high dimensions: Methods and scalable algorithms. 2023. URL https://arxiv.org/abs/2303.04258.

S. Ma, L. Xue, and H. Zou. Alternating direction methods for latent variable gaussian graphical model selection. *Neural Computation*, 25:2172–2198, 2012. URL https://direct.mit.edu/neco/article/25/8/2172/7900/Alternating-Direction-Methods-for-Latent-Variable.

N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006. URL https://projecteuclid.org/journals/annals-of-statistics/volume-34/issue-3/High-dimensional-graphs-and-variable-selection-with-the-Lasso/10.1214/009053606000000281.full.

I. Papastathopoulos and K. Strokorb. Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15, 2016. ISSN 0167-7152. URL https://www.sciencedirect.com/science/article/pii/S0167715215002874.

P. Ravikumar, M.J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2008. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-5/issue-none/High-dimensional-covariance-estimation-by-minimizing-%E2%84%931-penalized-log-determinant/10.1214/11-EJS631.full.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010. URL https://epubs.siam.org/doi/10.1137/070697835.

S. I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer, New York, 2008.

I. Echave-Sustaeta Rodríguez and F. Röttger. Latent gaussian graphical models with golazo penalty, 2024. URL https://arxiv.org/abs/2408.12482.

H. Rootzén and N. Tajvidi. Multivariate generalized Pareto distributions. *Bernoulli*, 12:917–930, 2006. URL https://projecteuclid.org/journals/bernoulli/volume-12/issue-5/Multivariate-generalized-Pareto-distributions/10.3150/bj/1161614952.full.

H. Rootzén, J. Segers, and J.L. Wadsworth. Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *J. Multivariate Anal.*, 165:117–131, 2018. URL https://www.sciencedirect.com/science/article/pii/S0047259X17303147.

F. Röttger, J.I. Coons, and A. Grosdos. Parametric and nonparametric symmetries in graphical models for extremes, 2023a. URL https://arxiv.org/abs/2306.00703.

F. Röttger, S. Engelke, and P. Zwiernik. Total positivity in multivariate extremes. *The Annals of Statistics*, 51(3):962 – 1004, 2023b. URL https://doi.org/10.1214/23-AOS2272.

J. Segers. One- versus multi-component regular variation and extremes of Markov trees. *Advances in Applied Probability*, 52:855–878, 2020.

A. Taeb and V. Chandrasekaran. Interpreting latent variables in factor models via convex optimization. *Mathematical Programming*, 167:129–154, 2016. URL https://link.springer.com/article/10.1007/s10107-017-1187-7.

A. Taeb, J.T. Reager, M.J. Turmon, and V. Chandrasekaran. A statistical graphical model of the California reservoir system. *Water Resources Research*, 53:9721 – 9739, 2017. URL https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017WR020412.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009. URL https://ieeexplore.ieee.org/document/4839045.

P. Wan and C. Zhou. Graphical lasso for extremes. 2023. URL https://arxiv.org/abs/2307.15004.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. URL https://www.jmlr.org/papers/volume7/zhao06a/zhao06a.pdf.

C. Zhou. Dependence structure of risk factors and diversification effects. *Risk and Insurance / Measures and Control 2*, 2009. URL https://api.semanticscholar.org/CorpusID:15682170.

J. Zscheischler and S. Seneviratne. Dependence of drivers affects risks associated with compound events. *Science Advances*, 3, 2017. URL https://www.science.org/doi/10.1126/sciadv.1700263.