

STATISTICAL INFERENCE FOR HÜSLER–REISS GRAPHICAL MODELS THROUGH MATRIX COMPLETIONS

Manuel Hentschel, Sebastian Engelke,
Johan Segers

LIDAM Discussion Paper ISBA
2022 / 32

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/isba/publication.html>

Statistical Inference for Hüsler–Reiss Graphical Models Through Matrix Completions

Manuel Hentschel

Research Center for Statistics, University of Geneva
and

Sebastian Engelke

Research Center for Statistics, University of Geneva
and

Johan Segers

ISBA/LIDAM, UCLouvain

October 27, 2022

Abstract

The severity of multivariate extreme events is driven by the dependence between the largest marginal observations. The Hüsler–Reiss distribution is a versatile model for this extremal dependence, and it is usually parameterized by a variogram matrix. In order to represent conditional independence relations and obtain sparse parameterizations, we introduce the novel Hüsler–Reiss precision matrix. Similarly to the Gaussian case, this matrix appears naturally in density representations of the Hüsler–Reiss Pareto distribution and encodes the extremal graphical structure through its zero pattern. For a given, arbitrary graph we prove the existence and uniqueness of the completion of a partially specified Hüsler–Reiss variogram matrix so that its precision matrix has zeros on non-edges in the graph. Using suitable estimators for the parameters on the edges, our theory provides the first consistent estimator of graph structured Hüsler–Reiss distributions. If the graph is unknown, our method can be combined with recent structure learning algorithms to jointly infer the graph and the corresponding parameter matrix. Based on our methodology, we propose new tools for statistical inference of sparse Hüsler–Reiss models and illustrate them on large flight delay data in the U.S.

Keywords: extreme value analysis; multivariate generalized Pareto distribution; sparsity; variogram

1 Introduction

In statistical modelling, conditional independence and graphical models are well-established concepts for analyzing structural relationships in data (Lauritzen, 1996; Wainwright and Jordan, 2008). Particularly important are Gaussian graphical models, also known as Gaussian Markov random fields (Rue and Held, 2005). The graphical structure of a multivariate normal distribution with positive definite covariance matrix Σ can be read off from the zeros of its precision matrix Σ^{-1} .

For risk assessment in fields such as climate science, hydrology, or finance, the primary interest is in extreme observations, with attention to both the marginal tails and the dependence between multiple risk factors. Multivariate extreme value theory provides asymptotically motivated models and statistical tools for extreme events. In view of the growing complexity and dimensionality of modern data sets, sparsity and graphical models are becoming crucial notions for the analysis of extremes (e.g., Engelke and Ivanovs, 2021).

There are two different ways of defining graphical models for extreme value distributions. The first is based on max-linear models (Gissibl and Klüppelberg, 2018) and the second one studies multivariate Pareto distributions (Engelke and Hitz, 2020). We follow the second approach since their new notions of conditional independence and extremal graphical models link naturally to an extremal version of the well-known Hammersley–Clifford theorem for density factorizations. Moreover, in the case of tree graphs, Segers (2020) shows that the extremes of regularly varying Markov trees converge to these extremal tree models. Lee and Joe (2018) propose parsimonious models for extreme value copulas; the link with extremal graphical models is made in Asenova et al. (2021).

The class of extremal graphical models with Hüsler–Reiss Pareto distributions is of particular interest. In d dimensions, the parameter of this family is a variogram matrix $\Gamma \in \mathbb{R}^{d \times d}$. Because of their flexibility and stochastic properties, Hüsler–Reiss distributions can be seen as the counterpart of the Gaussian family for multivariate extremes. In combination with extremal graphical models, the Hüsler–Reiss family constitutes a powerful tool for sparse extreme value modelling, with many open questions still to explore.

For a connected, undirected graph $G = (V, E)$ with nodes $V = \{1, \dots, d\}$ and edges E , Engelke and Hitz (2020) show that such a distribution’s graphical structure can be read off from a set of precision matrices $\Theta^{(k)} \in \mathbb{R}^{(d-1) \times (d-1)}$, for $k \in V$. The latter are defined as the inverses of the matrices obtained by the covariance mappings

(Farris et al., 1970)

$$\Sigma_{ij}^{(k)} = \frac{1}{2}(\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij}), \quad i, j \neq k.$$

While zeros in $\Theta^{(k)}$ correspond to extremal conditional independence of nodes $i, j \neq k$, the information on edges involving the k th node is encoded only indirectly through the row sums of this matrix. A natural question, appearing also in the discussion of Engelke and Hitz (2020), is if there exists a symmetric approach involving a single $d \times d$ precision matrix.

Statistical inference for Hüsler–Reiss graphical models is limited so far to the simple structures of trees and block graphs (Engelke and Volgushev, 2020; Asenova and Segers, 2021). The parameter matrix Γ is then additive on the graph and the maximum likelihood estimator is an explicit combination of the bivariate estimators on the edges. Since block graphs lack flexibility for general applications, several discussion contributions of Engelke and Hitz (2020) have emphasized the need for estimators suitable for more general graphs.

In this paper, we obtain new theoretical results on Hüsler–Reiss distributions that answer the two open questions above and enable statistical inference on extremal graphical models on decomposable and non-decomposable graphs. We first introduce the Hüsler–Reiss precision matrix $\Theta \in \mathbb{R}^{d \times d}$ as $\Theta_{ij} = \Theta_{ij}^{(k)}$ for some $k \neq i, j$, a definition which—surprisingly—is independent of the particular choice of $k \in V$. This positive semi-definite matrix indeed reflects the sparsity of the extremal graph by zero off-diagonal entries. We give several characterizations of this matrix, one of them as the Moore–Penrose inverse of a projection of the parameter matrix Γ .

For a given, general graph $G = (V, E)$, the new Hüsler–Reiss precision matrix allows us to represent the maximum (surrogate-)likelihood estimate of Γ on the graph G as the maximizer of the constrained optimization problem

$$\log |\Theta|_+ + \frac{1}{2} \text{tr}(\widehat{\Gamma}\Theta), \quad \text{s.t. } \Theta_{ij} = 0 \text{ if } (i, j) \notin E,$$

where $\widehat{\Gamma}$ is the empirical variogram (Engelke and Volgushev, 2020) and $|\cdot|_+$ denotes the pseudo-determinant. We prove that the solution to this optimization problem is given by the solution of a matrix completion problem. The aim of this completion is to find a conditionally negative definite variogram matrix Γ that has specified values in the entries corresponding to the edges E and zeros in the remaining entries of its precision matrix, i.e., $\Theta_{ij} = 0$ for $(i, j) \notin E$. For the case of decomposable graphs G , our completion algorithm is exact and has finitely many steps. For non-decomposable graphs, the solution is the limit of a converging sequence of variogram matrices. These results can be seen as a semi-definite extension of matrix completion problems for

the covariance matrix of Gaussian distributions studied in Speed and Kiiveri (1986) and Bakonyi and Woerdeman (2011).

Section 2 provides some preliminaries on extremal graphical models. The Hüsler–Reiss family is studied in Section 3, with a focus on the precision matrix Θ . Estimation of the Hüsler–Reiss variogram matrix of an extremal graphical model leads to matrix completion problems that are analyzed in Section 4. Statistical inference is treated in Section 5 while Section 6 reports on a case study involving data on delays in the domestic U.S. air travel network. The supplementary material contains mathematical details and proofs.

2 Extremal graphical models

2.1 Multivariate generalized Pareto distributions

Multivariate extreme value theory studies the tail behavior of a random vector $X = (X_1, \dots, X_d)$. A first summary of the extremal dependence structure of the bivariate margins for $i, j \in \{1, \dots, d\}$ is the extremal correlation $\chi_{ij} \in [0, 1]$, defined as

$$\chi_{ij} := \lim_{p \rightarrow 0} \chi_{ij}(p) := \lim_{p \rightarrow 0} \mathbb{P}(F_i(X_i) > 1 - p \mid F_j(X_j) > 1 - p), \quad (2.1)$$

whenever the limit exists and where F_i is the distribution function of X_i . When $\chi_{ij} > 0$ we say that X_i and X_j are asymptotically dependent, and when $\chi_{ij} = 0$ we speak of asymptotic independence. In the former case, there are two different, but closely related approaches for modelling extremal dependence: through component-wise maxima of independent copies of X leading to max-stable distributions (de Haan and Resnick, 1977); and through threshold exceedances of X resulting in multivariate generalized Pareto distributions (Rootzén and Tajvidi, 2006). Here, we concentrate on the threshold exceedance approach since it is well-suited for graphical modelling (Engelke and Hitz, 2020; Segers, 2020). For statistical models for asymptotic independence, we refer to Heffernan and Tawn (2004), for instance, and to Papastathopoulos et al. (2017) in the context of extremes of Markov chains.

To make abstraction of the univariate marginal distributions and concentrate on the extremal dependence, it is usually assumed that all variables X_i follow a given continuous distribution. Throughout, we use standard exponential margins, that is, $\mathbb{P}(X_i \leq x) = 1 - \exp(-x)$ for $x \geq 0$ and $i = 1, \dots, d$; we discuss this choice at the end of this section.

Let $\mathbf{0}$, $\mathbf{1}$, and ∞ denote the d -vectors with all elements equal to 0, 1, and ∞ respectively. A random vector $Y = (Y_1, \dots, Y_d)$ is said to follow a multivariate

generalized Pareto distribution (Rootzén and Tajvidi, 2006) if for any $z \in \mathcal{L} = \{x \in \mathbb{R}^d : x \not\leq \mathbf{0}\}$, we have

$$\mathbb{P}(Y \leq z) := \lim_{u \rightarrow \infty} \mathbb{P}(X - u\mathbf{1} \leq z \mid X \not\leq u) = \frac{\Lambda(z \wedge \mathbf{0}) - \Lambda(z)}{\Lambda(\mathbf{0})}, \quad (2.2)$$

for some vector X , which is then said to be in the domain of attraction of Y . The so-called exponent measure Λ is a measure on $[-\infty, \infty)^d \setminus \{-\infty\}$ that is finite on sets bounded away from $-\infty$, and $\Lambda(z) := \Lambda([-\infty, \infty)^d \setminus [-\infty, z])$. Multivariate generalized Pareto distributions are the only possible distributions that can arise as limits of threshold exceedances as in (2.2). We assume Λ to be absolutely continuous with respect to the d -dimensional Lebesgue measure and let λ denote its Radon–Nikodym derivative. The set of valid exponent measure densities λ is characterized by the following two properties:

$$\begin{aligned} \lambda(y + t\mathbf{1}) &= \exp(-t)\lambda(y), & \forall t \in \mathbb{R}, y \in \mathbb{R}^d, \\ \int_{y_i > 0} \lambda(y) dy &= 1, & \forall i \in \{1, \dots, d\}. \end{aligned}$$

Since the distribution of Y can be seen as the restriction of Λ to \mathcal{L} , its density f then also exists and is proportional to the exponent measure density λ as $f(y) = \lambda(y)/\Lambda(\mathbf{0})$ for all $y \in \mathcal{L}$. For a non-empty subset $I \subseteq \{1, \dots, d\}$, we define, with some abuse of terminology, the I -th marginal Y_I of Y as the distributional limit that arises in (2.2) when X is replaced by the sub-vector $X_I = (X_i)_{i \in I}$, also in the conditioning event. We have $f_I(y_I) = \lambda_I(y_I)/\Lambda_I(\mathbf{0})$ for all $y \in \mathcal{L}_I = \{x \in \mathbb{R}^{|I|} : x \not\leq \mathbf{0}\}$, where the objects f_I , λ_I , and Λ_I are defined analogously as for $I = \{1, \dots, d\}$. It is important to note that Y_I is not the actual marginal distribution of Y , since the latter would have support on the space $[-\infty, \infty)^{|I|} \setminus \{\infty\}$. Instead, with the definition above, Y_I is the marginal of Y that is conditioned to lie in the space \mathcal{L}_I . For further properties of multivariate generalized Pareto distributions, we refer to Rootzén et al. (2018) and Kiriliouk et al. (2018).

Remark 2.1. We consider multivariate generalized Pareto distributions Y with standard exponential marginals supported on $\mathcal{L}_Y = \{x \in \mathbb{R}^d : x \not\leq \mathbf{0}\}$ and exponent measure Λ_Y . Another common choice is the standard Pareto marginal distribution (e.g., Engelke and Hitz, 2020), in which case the multivariate Pareto distribution is that of $Z = \exp(Y)$, supported on $\mathcal{L}_Z = \{x \in [0, \infty)^d : x \not\leq \mathbf{1}\}$ with exponent measure $\Lambda_Z(x) = \Lambda_Y(\log x)$ and density

$$\lambda_Z(x) = \left(\prod_{i=1}^d x_i^{-1} \right) \lambda_Y(\log x), \quad x \in \mathcal{L}_Z.$$

All results in this paper are still applicable with these (or other) marginals since they only concern the dependence structure.

2.2 Extremal conditional independence

In classical statistics, graphical models are defined through conditional independence relations between components of a random vector. Since the support \mathcal{L} of a multivariate generalized Pareto distribution is not a product space, this definition is impractical in this case. Instead, Engelke and Hitz (2020) introduce a novel notion of extremal conditional independence in terms of the vectors $Y^{(k)} = (Y_1^{(k)}, \dots, Y_d^{(k)})$ for $k \in V := \{1, \dots, d\}$, defined as the multivariate generalized Pareto vector Y conditioned on the event $\{Y_k > 0\}$, and conveniently supported on the product spaces $\mathcal{L}^{(k)} = \{x \in \mathcal{L} : x_k > 0\}$. The extremal version of conditional independence between sub-vectors Y_A and Y_B given Y_C , for non-empty, disjoint sets $A \cup B \cup C = V$, is then defined as

$$\forall k \in V : Y_A^{(k)} \perp\!\!\!\perp Y_B^{(k)} \mid Y_C^{(k)}, \quad (2.3)$$

and it is denoted by $Y_A \perp_e Y_B \mid Y_C$; here, for $I \subseteq V$, we let $Y_I^{(k)} = (Y_i^{(k)})_{i \in I}$ denote the I -th marginal of $Y^{(k)}$ in the usual sense.

We consider the index set V as a set of nodes and let $E \subseteq \mathcal{E}(V)$ be the set of undirected edges of a connected graph $G = (V, E)$, where $\mathcal{E}(V)$ denotes the set of all possible edges. Figure 1 shows four different graph structures in increasing generality; see Section S.2 for definitions of the related terms. We say that Y follows an extremal graphical model on G if for $i, j \in V$ with $i \neq j$ we have

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

Engelke and Hitz (2020) show that the existence of a positive and continuous exponent measure density λ of Y implies the equivalence of (2.3) to the factorization

$$\lambda(y)\lambda_C(y_C) = \lambda_{AUC}(y_{AUC})\lambda_{BUC}(y_{BUC}), \quad y \in \mathcal{L}.$$

Moreover, if Y is an extremal graphical model on a decomposable graph G , then the density of Y factorizes on the graph into lower-dimensional marginal densities.

3 Hüsler–Reiss distributions

3.1 Definition

A popular class of multivariate generalized Pareto distributions are the Hüsler–Reiss distributions, which are parametrized by symmetric conditionally negative definite

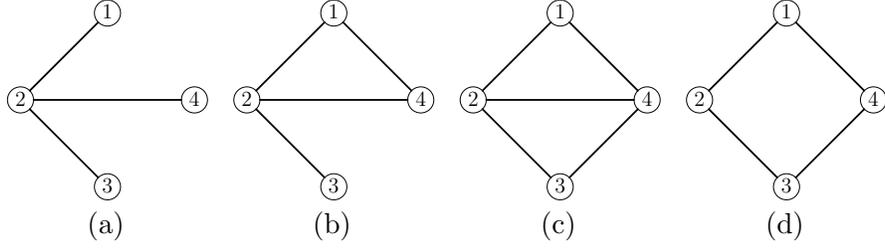


Figure 1: Different undirected, connected graphs on four nodes: (a) is a tree, (b) is a block graph, (c) is decomposable, (d) is non-decomposable.

matrices, here denoted by Γ (Hüsler and Reiss, 1989; Kabluchko et al., 2009). The set of these so-called variogram matrices is defined as

$$\mathcal{D}_d = \{\Gamma \in \mathbb{R}^{d \times d} : \Gamma = \Gamma^\top, \text{diag}(\Gamma) = \mathbf{0}, v^\top \Gamma v < 0 \forall \mathbf{0} \neq v \perp \mathbf{1}\}. \quad (3.1)$$

Variogram matrices are closely related to covariance matrices and can be constructed from a centered random vector W with covariance matrix Σ as

$$\Gamma_{ij} = \mathbb{E}(W_i - W_j)^2 = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}, \quad \forall i, j \in V.$$

Notably, this so-called covariance transform (Farris et al., 1970) is not injective and can be written for the entire matrix as

$$\gamma : \Sigma \mapsto \Gamma = \mathbf{1} \text{diag}(\Sigma)^\top + \text{diag}(\Sigma) \mathbf{1}^\top - 2\Sigma; \quad (3.2)$$

see Definition S.5.9 and Lemma S.5.10 for details.

Definition 3.1. Let $\Gamma \in \mathcal{D}_d$ be a variogram matrix, $k \in V$, and

$$\begin{aligned} \Sigma^{(k)} &= \varphi_k(\Gamma) := \frac{1}{2}(\Gamma_{ik} + \Gamma_{jk} - \Gamma_{ij})_{i,j \neq k} \in \mathbb{R}^{(d-1) \times (d-1)}, \\ \mu^{(k)} &= \left(-\frac{1}{2}\Gamma_{ik}\right)_{i \neq k} \in \mathbb{R}^{d-1}. \end{aligned}$$

A multivariate generalized Pareto random vector $Y = (Y_1, \dots, Y_d)$ is Hüsler–Reiss Pareto distributed with parameter matrix Γ if its exponent measure density λ satisfies

$$\lambda(y; \Gamma) = \frac{\exp(-y_k)}{\sqrt{(2\pi)^{d-1} |\Sigma^{(k)}|}} \exp\left(-\frac{1}{2} \|y_{V \setminus \{k\}} - \mathbf{1}y_k - \mu^{(k)}\|_{\Theta^{(k)}}^2\right), \quad y \in \mathbb{R}^d,$$

with $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ and using the notation $\|v\|_M^2 = v^\top M v$.

Notably, this expression does not depend on the choice of k (e.g., Engelke et al., 2015, Theorem 3.3).

3.2 Hüsler–Reiss precision matrix

An important property of Gaussian graphical models is that the conditional independence structure of a distribution with covariance matrix Σ can be read off from the zeros in the (Gaussian) precision matrix Σ^{-1} . Hüsler–Reiss graphical models satisfy a similar property. Indeed, let Y be a Hüsler–Reiss Pareto vector with parameter matrix $\Gamma \in \mathcal{D}_d$ that is an extremal graphical model on the undirected graph $G = (V, E)$. Engelke and Hitz (2020) show that the graph G is necessarily connected and that the graphical structure can be read off from the precision matrices $\Theta^{(k)} = (\Sigma^{(k)})^{-1}$ for any $k \in V$; in the following we index these $(d-1) \times (d-1)$ matrices by $V \setminus \{k\}$ for the sake of simpler notation. Two distinct nodes $i, j \neq k$ are extremal conditionally independent in the sense that $Y_i \perp_e Y_j \mid Y_{V \setminus \{i, j\}}$ if and only if the corresponding entry $\Theta_{ij}^{(k)}$ is zero. If one of the nodes, say j , is equal to k , extremal conditional independence is equivalent to the row sum $\sum_{l \neq k} \Theta_{il}^{(k)}$ being zero.

From this result it follows that each $\Theta^{(k)}$ contains all information on conditional independence of Y . A natural question, which was also raised in the discussion of Engelke and Hitz (2020), is whether there is a single $d \times d$ precision matrix that contains this information and is independent of a choice of k . In fact, such a matrix can be defined in the following way.

Definition 3.2. *Let $\Gamma \in \mathcal{D}_d$ be a variogram matrix. For $d \geq 3$ define $\Theta \in \mathbb{R}^{d \times d}$ by*

$$\Theta_{ij} = \Theta_{ij}^{(k)}, \quad \text{for some } k \in V \setminus \{i, j\},$$

with $\Theta^{(k)}$ as above. In the case $d = 2$, set $\Theta_{11} = \Theta_{22} = -\Theta_{12} = -\Theta_{21} = 1/\Gamma_{12}$.

Lemma 1 and Proposition 3 in Engelke and Hitz (2020) imply that the matrix Θ is well-defined and represents extremal conditional independence of nodes $i, j \in V$ in the corresponding Hüsler–Reiss model by

$$\Theta_{ij} = 0 \quad \iff \quad Y_i \perp_e Y_j \mid Y_{V \setminus \{i, j\}}. \quad (3.3)$$

While Definition 3.2 is a natural way to jointly represent the information contained in the $\Theta^{(k)}$ matrices, it remains to be shown that the matrix Θ allows for useful mathematical representations. In order to give a first such representation, let $(\cdot)^+$ denote the Moore–Penrose inverse (see Section S.1.1 for details), let I_d be the $d \times d$ identity matrix, write $\mathbf{e}_d = d^{-1}\mathbf{1}$, and let $\Pi = I_d - \mathbf{1}\mathbf{e}_d^\top$ be the $d \times d$ centering matrix, i.e., the projection matrix onto the orthogonal complement of $\mathbf{1}$. Recall γ from (3.2).

Proposition 3.3. *Let $\Gamma \in \mathcal{D}_d$ and $S \in \mathbb{R}^{d \times d}$ satisfying $\gamma(S) = \Gamma$. Then*

$$\Pi S \Pi = \Pi \left(-\frac{1}{2} \Gamma \right) \Pi. \quad (3.4)$$

Furthermore, the matrix Θ from Definition 3.2 satisfies

$$\Theta = (\Pi S \Pi)^+ \quad (3.5)$$

$$= \lim_{t \rightarrow \infty} (t \mathbf{1} \mathbf{1}^\top + S)^{-1}. \quad (3.6)$$

Since the mapping γ preserves (a)symmetry and Γ is symmetric, the condition $\gamma(S) = \Gamma$ requires S to be a symmetric matrix, as well. It turns out that, restricted to symmetric matrices, (3.4) is in fact equivalent to $\gamma(S) = \Gamma$; see Lemma S.5.8 for details.

In practice, the matrix S is usually a (definite or semi-definite) covariance matrix, so it is natural to only consider symmetric matrices here. For the sake of completeness we show in the proof of Proposition 3.3 that (3.4) is a sufficient and necessary condition for (3.5), also allowing asymmetric matrices S . Lemma S.5.7 shows that (3.4) is further equivalent to S being a generalized inverse of Θ in the sense $\Theta S \Theta = \Theta$.

In light of (3.4), an obvious choice of S for a given Γ is setting $S = -\frac{1}{2} \Gamma$, while other interesting choices for S are the matrices $\tilde{\Sigma}^{(k)}$ constructed in Engelke and Hitz (2020) by adding a k th row and column containing only zeros to the matrix $\Sigma^{(k)}$. Furthermore, these matrices relate to $\Pi(-\frac{1}{2} \Gamma) \Pi$ by $\Pi(-\frac{1}{2} \Gamma) \Pi = d^{-1} \sum_{k=1}^d \tilde{\Sigma}^{(k)} - t(\Gamma) \mathbf{1} \mathbf{1}^\top$ where $t(\Gamma) = \frac{1}{2} d^{-2} \mathbf{1}^\top \Gamma \mathbf{1}$ is the largest scalar such that $d^{-1} \sum_{k=1}^d \tilde{\Sigma}^{(k)} - t \mathbf{1} \mathbf{1}^\top$ is positive semi-definite.

To better understand representation (3.5), let $\mathcal{P}_d^1 \subset \mathbb{R}^{d \times d}$ denote the set of symmetric positive semi-definite matrices with kernel equal to $\text{span}(\{\mathbf{1}\})$. Further, let σ and θ be the mappings from a variogram Γ to the corresponding matrices Σ and Θ :

$$\begin{aligned} \sigma : \Gamma &\longmapsto \Sigma := \Pi \left(-\frac{1}{2} \Gamma \right) \Pi, \\ \theta : \Gamma &\longmapsto \Theta := \Sigma^+ = \sigma(\Gamma)^+ = \left(\Pi \left(-\frac{1}{2} \Gamma \right) \Pi \right)^+. \end{aligned} \quad (3.7)$$

Proposition 3.4. *The mappings σ and θ are homeomorphisms between \mathcal{D}_d and \mathcal{P}_d^1 with inverses*

$$\sigma^{-1}(\Sigma) = \gamma(\Sigma), \quad \theta^{-1}(\Theta) = \gamma(\Theta^+),$$

respectively, where γ denotes the covariance mapping defined in (3.2).

From this proposition it follows that the matrix Θ from Definition 3.2 is always positive semi-definite with kernel equal to $\text{span}(\{\mathbf{1}\})$. Furthermore, the class of Hüsler–Reiss distributions, parametrized by the collection \mathcal{D}_d of conditionally negative definite matrices, can just as well be parametrized by the set \mathcal{P}_d^1 , interpreted either in the role of Θ or Σ . As we will show in Corollary 3.7, the matrix Σ is the degenerate covariance matrix of a particular transformation of the Hüsler–Reiss Pareto vector Y . Similarly to the Gaussian case, the precision matrix Θ can be obtained from this covariance matrix as $\Theta = \Sigma^+$, using the Moore–Penrose inverse due to its singularity.

Representation (3.6) builds a bridge to the original work by Hüsler and Reiss (1989). They studied the asymptotic distribution of component-wise maxima of a Gaussian triangular array where each row is an independent random sample from a Gaussian distribution with $d \times d$ correlation matrix $\rho^{[n]}$ such that the limit

$$L := \lim_{n \rightarrow \infty} \log(n)(\mathbf{1}\mathbf{1}^\top - \rho^{[n]}) \quad (3.8)$$

exists. These component-wise maxima then converge to a max-stable Hüsler–Reiss distributions, parametrized by the conditionally negative definite matrix $\Gamma = 4L$; see (3.1) in Hüsler and Reiss (1989). If $S \in \mathbb{R}^{d \times d}$ is a matrix satisfying $\gamma(S) = \Gamma$ for a given $\Gamma \in \mathcal{D}_d$, and if $\rho^{[n]}$ is the correlation matrix associated to the covariance matrix

$$\Sigma^{[n]} = a_n \mathbf{1}\mathbf{1}^\top + S, \quad n \in \mathbb{N},$$

for a scalar sequence $a_n \rightarrow \infty$, then a straightforward calculation yields $a_n(\mathbf{1}\mathbf{1}^\top - \rho^{[n]}) \rightarrow \gamma(S)/2 = \Gamma/2$ as $n \rightarrow \infty$; see Lemma S.5.8 for the latter identity. Hence, choosing $a_n = 2 \log(n)$ yields (3.8) with $L = \Gamma/4$, as required. At the same time, Proposition 3.3 implies that the covariance matrices $\Sigma^{[n]}$ are eventually invertible and that the corresponding Gaussian precision matrices $\Theta^{[n]} = (\Sigma^{[n]})^{-1}$ converge to the Hüsler–Reiss precision matrix $\Theta = \theta(\Gamma)$ as $n \rightarrow \infty$.

Remark 3.5. The Hüsler–Reiss precision matrix Θ is motivated by its connection to extremal conditional independence in (3.3). It turns out that this matrix is also useful to describe other stochastic properties of the Hüsler–Reiss distributions. Based on the results of the present paper, Röttger et al. (2021) show that multivariate total positivity of order two, a notion of positive dependence, can be encoded as $\Theta_{ij} \leq 0$ for all $i \neq j$, and Engelke et al. (2022b) suggest that the precision matrix can be used to estimate extremal graphical structures by penalizing its L^1 norm $\|\Theta\|_1$. Compared to the Gaussian case, a difficulty of dealing with Θ in mathematical derivations and statistical implementations is that it is positive semi-definite and not invertible.

3.3 Hüsler–Reiss exponent measure density

For a given Γ , the characterizations in Proposition 3.3 give a straightforward way to compute Θ in Definition 3.2 as $\Theta = \theta(\Gamma) = (\Pi(-\frac{1}{2}\Gamma)\Pi)^+$ and thus retrieve the conditional independence structure of the corresponding extremal graphical model. In a Gaussian model, another property of the precision matrix is that it can be used to express its probability density function in a concise way. The following result shows that similar expressions are also possible for Hüsler–Reiss distributions. Recall the notation $\mathbf{e}_d = d^{-1}\mathbf{1}$ and $\|v\|_M = \sqrt{v^\top M v}$ for $v \in \mathbb{R}^d$ and positive semi-definite $M \in \mathbb{R}^{d \times d}$.

Proposition 3.6. *Let $\Gamma \in \mathcal{D}_d$ and $\Theta = \theta(\Gamma)$. Then the exponent measure density $\lambda(\cdot; \Gamma)$ from Definition 3.1 can be expressed as*

$$\lambda(y; \Theta) = c_\Theta \cdot \exp\left(-\frac{1}{2}y^\top \Theta y - y^\top (\mathbf{e}_d - r_\Theta)\right), \quad y \in \mathbb{R}^d, \quad (3.9)$$

with $r_\Theta = \Theta(-\frac{1}{2}\Gamma)\mathbf{e}_d$. Furthermore, let $v \in \mathbb{R}^d$ be such that $v^\top \mathbf{1} = 1$ and let $\mu_v = \Theta^+(r_\Theta + v) = \Pi(-\frac{1}{2}\Gamma)v$. Then λ can also be written as

$$\lambda(y; \Theta) = c_{\Theta, v} \cdot \exp(-v^\top y) \cdot \exp\left(-\frac{1}{2}\|y - \mu_v\|_\Theta^2\right). \quad (3.10)$$

The constants c_Θ and $c_{\Theta, v}$ are such that $\int_{y_i > 0} \lambda(y) dy = 1$ for all $i \in V$, and are stated explicitly in the proof.

Note that the value of the right-hand side in (3.10) does not depend on $v \in \mathbb{R}^d$, as long as $v^\top \mathbf{1} = 1$, which is why v is omitted on the left-hand side. In line with the stochastic representation given in expression (28) of Engelke and Hitz (2020), consider the following construction. For $\Theta \in \mathcal{P}_d^1$ and for $\mathbf{0} \leq v \in \mathbb{R}^d$ satisfying $v^\top \mathbf{1} = 1$, let μ_v be as in Proposition 3.6, define $\Pi_v = I_d - \mathbf{1}v^\top$, and let $W_v \sim \mathcal{N}(\Pi_v \mu_v, \Pi_v \Theta^+ \Pi_v^\top)$, be a d -dimensional random vector with degenerate normal distribution on $\{v\}^\perp$. For a standard exponential random variable $R \sim \text{Exp}(1)$ put

$$Y_v = W_v + R \cdot \mathbf{1}. \quad (3.11)$$

Figure 2 illustrates this construction in \mathbb{R}^2 ; for a more detailed explanation see the proof of Corollary 3.7. Comparing the density of Y_v to (3.10) yields the following result. Note that the condition that $v \geq \mathbf{0}$ implies that $\{y \in \mathbb{R}^d : v^\top y > 0\}$ is contained in $\{y \in \mathbb{R}^d : y \not\leq \mathbf{0}\}$, the support of a multivariate generalized Pareto distribution.

Corollary 3.7. *Let Θ , v , and Y_v be as above. Let Y be a Hüsler–Reiss Pareto random vector with variogram matrix $\Gamma = \theta^{-1}(\Theta) = \gamma(\Theta^+)$. Then Y conditioned on the event $\{v^\top Y > 0\}$ is equal in distribution to Y_v .*

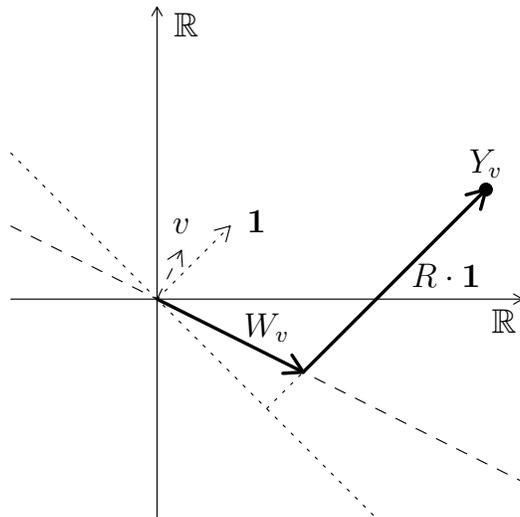


Figure 2: Geometric illustration of the stochastic representation of the conditioned Hüsler–Reiss distribution Y_v in Corollary 3.7. The random vector Y_v is decomposed into W_v on the hyperplane $\{v\}^\perp$ (dashed line) and radial part $R \cdot \mathbf{1}$, which is orthogonal to the hyperplane $\{\mathbf{1}\}^\perp$ (dotted line).

The results presented so far show that our definition of the Hüsler–Reiss precision matrix is in fact a very natural one. Similar to the Gaussian precision matrix it encodes the conditional independence structure of the corresponding graphical model, and the corresponding exponent measure density can be represented using Θ in a similar fashion to the Gaussian density. Corollary 3.7 shows that this similarity is due to the fact that a Hüsler–Reiss Pareto random vector can be decomposed into a linear transformation of a degenerate normal part with covariance matrix $\Sigma = \Theta^+$ and an exponentially distributed part which is identical for all possible Hüsler–Reiss Pareto distributions.

Using the general representation in (3.10), we recover different forms of the Hüsler–Reiss density from the literature. Setting $v = e_k$ to the k th unit vector for some $k \in V$, yields the density representation in Definition 3.1. The corresponding random vectors $Y^{(k)} = Y_{e_k}$ play a crucial role in statistical inference (Engelke et al., 2015) and exact simulation (Dombry et al., 2016). A careful look at other characterizations such as in Wadsworth and Tawn (2014), or the slightly more general definitions of the Hüsler–Reiss exponent measure density in Ho and Dombry (2019) and Kiriliouk et al. (2018), shows that our precision matrix Θ appears naturally in these parameterizations. Section S.3 makes these connections precise.

4 Matrix completion problems

A well-studied problem related to Gaussian graphical models is the task of constructing a model with a given graphical structure and specified marginal distributions on the fully connected subsets of vertices, or equivalently, to complete a partially defined covariance matrix such that its precision matrix has zeros where there are no edges in the specified graph (see Speed and Kiiveri, 1986; Bakonyi and Woerdeman, 2011). For extremal Hüsler–Reiss models, the same problem can be posed, and expressed as a matrix completion problem on the variogram and precision matrix.

To formalize the notion of a partially specified matrix, we define the set $\mathring{\mathbb{R}} := \mathbb{R} \cup \{“?”\}$, consisting of the real numbers and the placeholder “?” for unspecified values (see e.g., Bakonyi and Woerdeman, 2011, for this use of “?”). For an undirected graph $G = (V, E)$, a matrix is said to be “specified on G ” if it is specified on the diagonal and the entries corresponding to the edges of G . A matrix is said to be “partially conditionally negative definite” if it is symmetric, its diagonal is fully specified, and all fully specified principal submatrices are conditionally negative definite (here, a principal submatrix is any submatrix obtained by removing the same index set from both the columns and rows of the matrix). In computations involving partially specified matrices, an entry in the result is “?” as soon as any of the entries used to compute it is itself “?”.

Definition 4.1. *Let $G = (V, E)$ be an undirected graph and let $\mathring{\Gamma} \in \mathring{\mathbb{R}}^{d \times d}$ be a partially conditionally negative definite matrix, specified on G . The corresponding matrix completion problem is to find a conditionally negative definite matrix $\Gamma \in \mathcal{D}_d$ and $\Theta = \theta(\Gamma)$ such that*

$$\begin{aligned} \Gamma_{ij} &= \mathring{\Gamma}_{ij} & \forall (i, j) \in \overline{E}, \\ \Theta_{ij} &= 0 & \forall (i, j) \notin \overline{E}, \end{aligned} \tag{4.1}$$

where \overline{E} denotes the edge set E augmented by the diagonal entries $\{(i, i) : i \in V\}$.

An example for such a matrix completion problem on the graph in Figure 1b is given by

$$\mathring{\Gamma} = \begin{pmatrix} 0 & 3 & ? & 1 \\ 3 & 0 & 10 & 2 \\ ? & 10 & 0 & ? \\ 1 & 2 & ? & 0 \end{pmatrix}, \quad \Theta = \begin{pmatrix} ? & ? & 0 & ? \\ ? & ? & ? & ? \\ 0 & ? & ? & 0 \\ ? & ? & 0 & ? \end{pmatrix}. \tag{4.2}$$

In the Gaussian case, a similar problem can be posed with the covariance matrix Σ instead of Γ and with $\Theta = \Sigma^{-1}$. To this positive definite completion problem, an

explicit solution for decomposable graphs and a convergent algorithm for general graphs is given in Speed and Kiiveri (1986). In this section we discuss semi-definite matrix completion problems for Hüsler–Reiss models as in Definition 4.1, starting from simple graph structures such as trees and finishing with general graphs. We assume throughout that the graph G is connected since only those can be associated to non-degenerate Hüsler–Reiss models; see Section 3.2.

4.1 Trees and block graphs

Block graphs are simple graph structures where the separator sets contain only single nodes. Trees are a special case of this class where all cliques consist of exactly two nodes. The particular structure of block graphs makes them appealing for the construction of parametric models and for statistical inference (Engelke and Hitz, 2020; Asenova et al., 2021). In fact, this structure also yields a simple explicit solution to the matrix completion problem in Definition 4.1. If $\mathring{\Gamma}$ is a partially specified matrix on the connected block graph $G = (V, E)$, then a unique completion exists (Engelke and Hitz, 2020, Prop. 4) and can be expressed as

$$\Gamma_{ij} = \sum_{(s,t) \in \text{path}(i,j)} \mathring{\Gamma}_{st}, \quad (4.3)$$

where $\text{path}(i, j)$ is the unique shortest path between i and j in G ; see also Engelke and Volgushev (2020) and Asenova and Segers (2021). Using this result, the missing entries in (4.2) can be computed to be $\Gamma_{13} = \Gamma_{31} = 13$ and $\Gamma_{34} = \Gamma_{43} = 12$.

4.2 Decomposable graphs

In this section, a solution to the matrix completion problem will be given for connected, decomposable graphs. First, consider the simplest (non-trivial) example from this class of graphs, a graph consisting of exactly two cliques. Recall the definitions of γ from (3.2) and φ_k from Definition 3.1, and observe that its inverse can be expressed as $\varphi_k^{-1}(\Sigma^{(k)}) = \gamma(\tilde{\Sigma}^{(k)})$, where $\tilde{\Sigma}^{(k)}$ is identical to $\Sigma^{(k)}$, with an additional k th row and column of zeros.

Lemma 4.2. *Let $G = (V, E)$ be a connected decomposable graph consisting of two cliques C_1, C_2 , separated by $D_2 = C_1 \cap C_2 \neq \emptyset$. Let $\mathring{\Gamma}$ be a partially conditionally negative definite matrix, specified on G . For some $k \in D_2$, let $\mathring{\Sigma}^{(k)} = \varphi_k(\mathring{\Gamma})$ and let $\Sigma^{(k)}$ be its unique positive definite completion with graphical structure $G|_{V \setminus \{k\}}$. Then*

$$\Gamma := \varphi_k^{-1}(\Sigma^{(k)}) = \gamma(\tilde{\Sigma}^{(k)}),$$

is the unique solution of the matrix completion problem 4.1 for $\mathring{\Gamma}$ and graph G .

Using this result, a completion for a general decomposable graph can be constructed by ordering its cliques C_1, \dots, C_m according to the running intersection property (see Section S.2) and iteratively applying Lemma 4.2 to the graphs $G_i = (V_i, E_i)$, $i = 2, \dots, m$, with $V_i = C_1 \cup \dots \cup C_i$, and edge sets $E_i = (C_i \times C_i) \cup (C' \times C')$, where $C' = C_1 \cup \dots \cup C_{i-1}$.

Proposition 4.3. *Let $G = (V, E)$ be a connected, decomposable graph and $\mathring{\Gamma}$ a partially conditionally negative definite matrix, specified on G . Then there exists a unique solution to the matrix completion problem 4.1 with $\mathring{\Gamma}$ and G . This solution can be computed explicitly as described above.*

The class of decomposable graphs is a significant extension of the class of block graphs. For instance, most graphs can be approximated in a non-trivial fashion by using their so-called decomposable completion. In contrast, the only block graph completion of a biconnected graph (i.e., a graph that remains connected after removal of any one vertex) is already the complete graph, since the only biconnected block graph is the complete graph, and adding edges to a graph preserves connectivity.

Example 4.4. Figure 3 shows an example of the matrix completion algorithm from Proposition 4.3. Starting with the clique $\{1, 2, 3\}$, the missing values are computed clique by clique. Edges whose corresponding matrix entries were already computed, are considered as part of the graph in subsequent steps, such that each computation is a direct application of Lemma 4.2. Thanks to the running intersection property, the required conditional independence structure is preserved.

4.3 General graphs

The class of general connected graphs is much larger than the class of decomposable ones; see Figure 1d for an example. In applications, when the graph is estimated from data without restrictions, non-decomposable structures often arise. The following results provide a more general but slightly weaker solution to the matrix completion problem in Definition 4.1, where, as before, \overline{E} is equal to $E \cup \{(i, i) : i \in V\}$.

Proposition 4.5. *Let $G = (V, E)$ be a connected graph and let $\mathring{\Gamma} \in \mathbb{R}^{d \times d}$ be specified on G , such that there exists a fully specified conditionally negative definite matrix that agrees with $\mathring{\Gamma}$ on the entries $(i, j) \in \overline{E}$. Then there exists a unique conditionally negative definite matrix Γ that solves the matrix completion problem (4.1) for $\mathring{\Gamma}$ and G .*

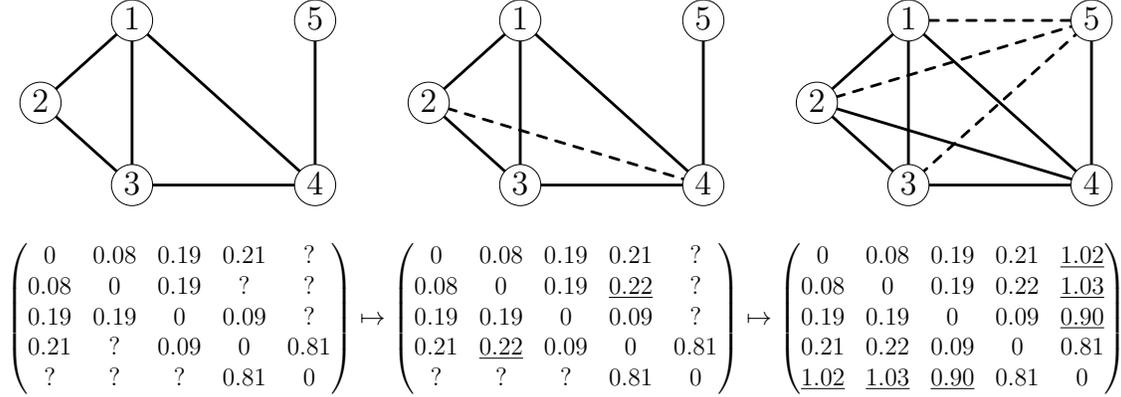


Figure 3: Illustration of Example 4.4. On the left-hand side is the initial partial matrix $\overset{\circ}{\Gamma}$ and the corresponding decomposable graph. In each of the following steps, dashed edges (top) correspond to newly computed matrix entries (underlined, bottom).

This result provides the same theoretical existence of a unique solution as Proposition 4.3 and allows the definition of the following mapping. We use the notation $\Gamma|_G \in \mathbb{R}^{d \times d}$ to denote the restriction of a fully specified matrix Γ to the entries corresponding to a graph $G = (V, E)$, in the sense

$$(\Gamma|_G)_{ij} = \begin{cases} \Gamma_{ij} & (i, j) \in \overline{E}, \\ \text{"?"} & (i, j) \notin \overline{E}. \end{cases} \quad (4.4)$$

Definition 4.6. For a connected graph $G = (V, E)$ let $\overset{\circ}{\mathcal{D}}_G = \{(\Gamma'|_G) : \Gamma' \in \mathcal{D}_d\}$ be the restriction of conditionally negative definite matrices to G . The function

$$\mathfrak{C}_G : \overset{\circ}{\mathcal{D}}_G \longrightarrow \mathcal{D}_d, \quad \overset{\circ}{\Gamma} \longmapsto \Gamma,$$

maps a partial matrix $\overset{\circ}{\Gamma} \in \overset{\circ}{\mathcal{D}}_G$ to its unique completion Γ satisfying the matrix completion problem in Definition 4.1 with respect to G .

Notably, the existence of any conditionally negative definite completion of $\overset{\circ}{\Gamma}$ is sufficient for the existence of a graphical completion. In the decomposable case, this can be guaranteed by verifying the definiteness of all fully specified principal submatrices, but in the general case this criterion does not work, as the following counter-example shows.

Example 4.7. For $d \geq 4$, let $\overset{\circ}{\Gamma} \in \mathbb{R}^{d \times d}$ be a partial matrix on the d -dimensional ring graph with entries $\overset{\circ}{\Gamma}_{ij} = 1$ if $|i - j| = 1$, $\overset{\circ}{\Gamma}_{1d} = \overset{\circ}{\Gamma}_{d1} = (2d)^2$, zeros on the diagonal, and

“?” elsewhere. Then all fully specified principal submatrices of this $\overset{\circ}{\Gamma}$ are conditionally negative definite, but there exists no conditionally negative definite completion of the entire matrix (see Section S.5.4.3 for details).

In general, computing the completion is less straightforward than in the decomposable case, but the following algorithm follows from the proof of Proposition 4.5.

Corollary 4.8. *Let G and $\overset{\circ}{\Gamma}$ be as in Proposition 4.5, and let $\Gamma_0 \in \mathcal{D}$ be such that $(\Gamma_0)|_G = \overset{\circ}{\Gamma}$. For some fixed $m \in \mathbb{N}$, let $G_i = (V, E_i)$, $i = 1, \dots, m$, be a set of decomposable graphs, such that $\bigcap_i E_i = E$ and define $(\Gamma_n)_{n \geq 1}$ recursively by*

$$\Gamma_n = \mathfrak{C}_{G_t} \left((\Gamma_{n-1})|_{G_t} \right),$$

with $t = t_n \equiv n \pmod{m}$ and where \mathfrak{C}_{G_t} is computed as in Proposition 4.3. Then Γ_n converges to the unique completion $\mathfrak{C}_G(\overset{\circ}{\Gamma})$ as $n \rightarrow \infty$.

Remark 4.9. An easy way to construct a suitable set $\{G_1, \dots, G_m\}$ is to set $\{e_1, \dots, e_m\} = \mathcal{E}(V) \setminus E$ and use $E_i = \mathcal{E}(V) \setminus \{e_i\}$, where $\mathcal{E}(V)$ denotes the set of all possible edges, see Section S.2. However, in practice this choice leads to a very slow convergence since each iteration only updates one entry in Γ and a $(d-3) \times (d-3)$ -dimensional matrix needs to be inverted to do so. Better performance can be achieved by choosing the G_i to be decomposable completions of G , with separator sets being as small as possible; see Baz et al. (2022) for details and further optimizations.

In order to apply Corollary 4.8, an initial (non-graphical) completion of $\overset{\circ}{\Gamma}$, to be used as Γ_0 , is required. The problem of finding such a matrix is also known as the Euclidean distance matrix completion problem, with solution algorithms for example in Bakonyi and Johnson (1995) and Fang and O’Leary (2012).

Example 4.10. To illustrate the algorithm from Corollary 4.8, consider the matrix Γ_0 below, which does not have any non-trivial graphical structure, and its “completion” Γ , whose conditional independence structure is described by the graph in Figure 4.

$$\Gamma_0 = \begin{pmatrix} 0 & 0.23 & \underline{0.08} & 0.09 & 0.21 \\ 0.23 & 0 & 0.14 & \underline{0.23} & \underline{0.19} \\ \underline{0.08} & 0.14 & 0 & 0.11 & \underline{0.20} \\ 0.09 & \underline{0.23} & 0.11 & 0 & 0.16 \\ 0.21 & \underline{0.19} & \underline{0.20} & 0.16 & 0 \end{pmatrix} \mapsto \Gamma = \begin{pmatrix} 0 & 0.23 & \underline{0.17} & 0.09 & 0.21 \\ 0.23 & 0 & 0.14 & \underline{0.20} & \underline{0.35} \\ \underline{0.17} & 0.14 & 0 & 0.11 & \underline{0.26} \\ 0.09 & \underline{0.20} & 0.11 & 0 & 0.16 \\ 0.21 & \underline{0.35} & \underline{0.26} & 0.16 & 0 \end{pmatrix}$$

These two matrices differ only in the highlighted entries, corresponding to non-edges in G . During the computation of Γ , the entries of Γ_0 that correspond to edges in G

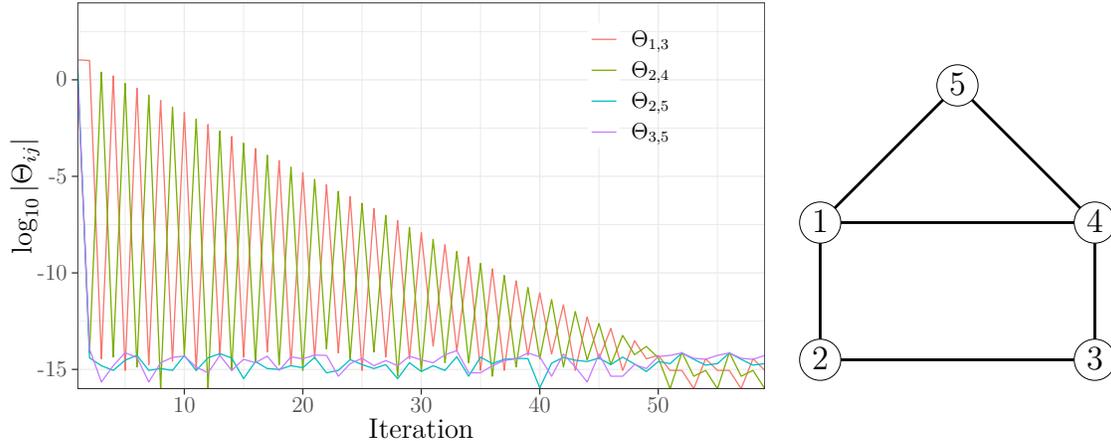


Figure 4: Illustration of Proposition 4.5 in Example 4.10. A non-decomposable graph $G = (V, E)$ (right), and the convergence of Θ_{ij} to zero for $(i, j) \notin E$ as the graphical completion of Γ is computed (left). In each iteration, the graph defined by edge set $E_1 = E \cup \{(1, 3)\}$ or $E_2 = E \cup \{(2, 4)\}$ is completed using Proposition 4.3.

do not change, and the convergence of the remaining entries in Θ to zero is plotted in Figure 4. It can be seen that the entries corresponding to edges $(2, 5)$ and $(3, 5)$ stay at zero (up to numerical precision of magnitude 10^{-15}) after the first few iterations, whereas the maximum of entries $(2, 4)$ and $(1, 3)$ converges to zero at a slower rate. This observation is in line with the fact that vertices $\{1, 2, 3, 4\}$ induce the chordless cycle that makes the graph non-decomposable.

5 Statistical inference

Estimation of a Hüsler–Reiss parameter matrix Γ that is an extremal graphical model on a given, known graph $G = (V, E)$ is currently restricted to the simple structures of trees (Engelke and Volgushev, 2020; Hu et al., 2022), latent trees (Asenova et al., 2021) and block graphs (Engelke and Hitz, 2020; Asenova and Segers, 2021). Since many data sets require more general graph structures, this has been pointed out as an important drawback for statistical modelling in the discussion of Engelke and Hitz (2020).

In this section we solve this issue by applying our results on matrix completions for variograms. In particular, we show how any consistent estimator $\widehat{\Gamma}$ can be transformed

into a consistent estimator $\widehat{\Gamma}^G$ with desired extremal graph structure. When the graph is unknown, structure learning methods \widehat{G} exist that recover the underlying graph consistently. Our completion then produces the first estimator $\widehat{\Gamma}^{\widehat{G}}$ that jointly estimates the graph structure and the parameter matrix for general graphs G in a consistent way.

5.1 Matrix completion as likelihood optimization

For a more statistical perspective on the matrix completions considered in the previous section, we first characterize them as constrained maximum likelihood estimators; see Uhler (2017) for the Gaussian case. For a variogram matrix $\bar{\Gamma} \in \mathcal{D}_d$ and a connected graph $G = (V, E)$, we will show that the maximizer of the positive semi-definite Gaussian log-likelihood

$$\log |\Theta|_+ + \frac{1}{2} \text{tr}(\bar{\Gamma}\Theta), \quad (5.1)$$

under suitable graph constraints, is equal to our completion operator \mathfrak{C}_G from Definition 4.6 applied to $\bar{\Gamma}|_G$. The connection to likelihood estimation for Hüsler–Reiss distributions arises by choosing for $\bar{\Gamma}$ the empirical variogram $\widehat{\Gamma}$ (Engelke and Volgushev, 2020). In this case, (5.1) is the (surrogate) log-likelihood of the Hüsler–Reiss model parameterized in terms of the precision matrix Θ (Röttger et al., 2021, Section 5.1). Section S.4.1 provides a simple derivation of this. Recall the notation $\bar{E} = E \cup \{(i, i) : i \in V\}$ for $E \subseteq \mathcal{E}(V)$ as well as the map θ and its inverse in Proposition 3.4.

Proposition 5.1. *Let $\bar{\Gamma} \in \mathcal{D}_d$ be a variogram matrix and $G = (V, E)$ a connected graph. Then $\mathfrak{C}_G(\bar{\Gamma}|_G) = \theta^{-1}(\bar{\Theta}_G)$ where $\bar{\Theta}_G$ is the unique maximizer of (5.1) over all Hüsler–Reiss precision matrices $\Theta \in \mathcal{P}_d^1$ under the constraint*

$$\Theta_{ij} = 0, \quad \forall (i, j) \notin \bar{E}.$$

We note that the result holds for any variogram matrix $\bar{\Gamma}$, but only for the empirical variogram $\widehat{\Gamma}$ there is an interpretation in terms of maximum (surrogate) likelihood estimation. In practice, solving this optimization provides an alternative to Corollary 4.8 to compute the graphical completion of a matrix. We leave this approach for future research.

5.2 Consistency

In order to show consistency results, a useful property of the completion \mathfrak{C}_G is that it is a continuous mapping from the space $\mathring{\mathcal{D}}_G$ of partially specified variogram matrices

to the space of variogram matrices \mathcal{D} .

Lemma 5.2. *The mapping \mathfrak{C}_G from Definition 4.6 is continuous.*

This continuity implies that a consistent estimator on the edges of a Hüsler–Reiss graphical model can be extended to a consistent estimator of the whole parameter matrix.

Theorem 5.3. *Consider a Hüsler–Reiss graphical model with graphical structure $G = (V, E)$ and variogram matrix Γ . Let $\hat{\Gamma}$ be an estimator for $\Gamma|_G$, which satisfies $\text{diag}(\hat{\Gamma}) \equiv \mathbf{0}$, is symmetric, and is consistent in the sense that for all $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\max_{(i,j) \in E} \left| \hat{\Gamma}_{ij} - \Gamma_{ij} \right| < \varepsilon\right) \rightarrow 1, \quad n \rightarrow \infty. \quad (5.2)$$

Then with probability tending to one there exists a completion of $\hat{\Gamma}$, that is, $\mathbb{P}(\hat{\Gamma} \in \mathring{\mathcal{D}}_G) \rightarrow 1$ as $n \rightarrow \infty$. Let $\hat{\Gamma}^G = \mathfrak{C}_G(\hat{\Gamma})$ denote this completion and $\hat{\Theta}^G$ the corresponding precision matrix (and set both matrices to the zero matrix if the completion does not yet exist). Then $\hat{\Gamma}^G$ is a consistent estimator for Γ with the correct graph structure, that is, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{(i,j) \in V \times V} \left| \hat{\Gamma}_{ij}^G - \Gamma_{ij} \right| < \varepsilon\right) \rightarrow 1, \quad n \rightarrow \infty,$$

and $\hat{\Theta}_{ij}^G = 0$ if $(i, j) \notin \bar{E}$.

This result provides the first consistent method to estimate the parameters of a Hüsler–Reiss distribution on a general graph $G = (V, E)$. Indeed, the only ingredient that is needed is a consistent estimator of Γ on the edge set E . There are many different possibilities for such estimators in the literature. A natural and computationally efficient estimator is the empirical variogram (Engelke et al., 2015; Engelke and Volgushev, 2020), which for the Hüsler–Reiss distributions is the empirical version of the parameter matrix Γ . Other proposals include M-estimators (Einmahl et al., 2012, 2016; Lalancette et al., 2021), proper scoring rules (de Fondeville and Davison, 2018) and likelihood methods. The latter have the advantage that they can incorporate censoring of components of the data that are not extreme (Ledford and Tawn, 1996; Wadsworth and Tawn, 2014). Pairwise likelihood methods (Padoan et al., 2010) reduce the high computational cost of censoring.

In order to guarantee that even for a fixed sample size n there exists a completion, it can be advantageous to start with a consistent estimator $\widehat{\Gamma} \in \mathcal{D}$ of the entire parameter matrix Γ . Since $\widehat{\Gamma}|_G \in \mathring{\mathcal{D}}_G$ by definition, Proposition 4.5 ensures that

$$\widehat{\Gamma}^G = \mathfrak{C}_G(\widehat{\Gamma}|_G) \quad (5.3)$$

exists and is a consistent estimator with graph structure G . In larger dimensions $d = |V|$, estimating all entries of Γ is often infeasible for any estimator that involves costly optimization or censoring, and the only option is then the empirical variogram or estimators based on simple summary statistics such as extremal coefficients (Einmahl et al., 2018). If the graph G is sparse, then there is an efficient alternative to estimating every entry of Γ , which requires only estimation on all cliques separately; see Section S.4.2.

The graph G is in practice often unknown and has to be estimated from the data. Consistent structure estimation methods for extremal graphs exist for trees (Engelke and Volgushev, 2020; Hu et al., 2022), and for general graphs based on lasso-type L^1 penalization (Engelke et al., 2022b). The latter paper proposes the `EGlearn` method and shows, under certain conditions, its sparsistency even in the high-dimensional case where the dimension may grow with the sample size. More precisely, if G is the graph implied by the zero entries in the true Θ , and $\widehat{G} = (V, \widehat{E})$ is the estimated graph from `EGlearn` then

$$\mathbb{P}(\widehat{G} = G) \rightarrow 1, \quad n \rightarrow \infty. \quad (5.4)$$

While they consistently recover a general graph structure, they do not obtain an estimate of the Γ matrix on the estimated graph structure.

Our theory complements the structure estimation in Engelke et al. (2022b). In combination, we are now able to estimate jointly any graph structure and the corresponding Hüsler–Reiss parameter matrix consistently.

Corollary 5.4. *Consider a Hüsler–Reiss graphical model with graphical structure $G = (V, E)$ and variogram matrix Γ . Let $\widehat{\Gamma}$ be a consistent estimator for Γ and $\widehat{G} = (V, \widehat{E})$ a sparsistent estimator of G as in (5.4). If $\widehat{\Gamma}^{\widehat{G}} = \mathfrak{C}_{\widehat{G}}(\widehat{\Gamma}|_{\widehat{G}})$ is the completion on \widehat{G} and $\widehat{\Theta}^{\widehat{G}}$ its precision matrix, then $\widehat{\Gamma}^{\widehat{G}}$ is a consistent and sparsistent estimator for Γ , that is, for all $\varepsilon > 0$,*

$$\mathbb{P}\left(\max_{(i,j) \in V \times V} \left| \widehat{\Gamma}_{ij}^{\widehat{G}} - \Gamma_{ij} \right| < \varepsilon, \quad \widehat{\Theta}_{ij}^{\widehat{G}} = 0 \quad \forall (i,j) \notin \overline{E}\right) \rightarrow 1, \quad n \rightarrow \infty.$$

The first statement in the above probability expression guarantees that $\widehat{\Gamma}^{\widehat{G}}$ is a consistent estimator of Γ . The second statement guarantees that, with probability going to one, the estimator $\widehat{\Gamma}^{\widehat{G}}$ is a graphical model with respect to the true graphical structure G .

6 Application

Extreme value analysis is often used to evaluate the risk of climate extremes such as heatwaves (e.g., Reich et al., 2014) or floods (e.g., Asadi et al., 2015; Cooley and Thibaud, 2019), and our methods are perfectly suitable for such data. In order to extend the range of possible applications, we illustrate our methodology on a different type of data, namely large flight delays in the U.S. flight network. Excessive delays have a variety of negative effects, ranging from inconveniences for passengers and congestion of critical airport infrastructure to financial losses for involved parties.

6.1 Data

The United States Bureau of Transportation Statistics¹ provides records of domestic flights in the U.S. that are operated by major carriers (at least 1% market share) at airports accounting for at least 1% of domestic enplanements. We use this data from 2005 to 2020, and filter it by selecting only airports from the contiguous U.S. with a minimum of 1000 flights per year. For each airport, we compute the accumulated (positive) flight delays (in minutes) on a given day. Considering that there are a few days for which no data are available, this results in a data set with $n = 5601$ observations $x_1, \dots, x_n \in \mathbb{R}^d$ of daily accumulated flight delays for the 16 years at $d = 170$ airports. This pre-processed data set is available in the R-package `graphicalExtremes` (Engelke et al., 2022a).

6.2 Exploratory analysis

The models studied in this paper are suitable for variables where the largest observations are asymptotically dependent. For our data set, such dependence between the largest daily accumulated flight delays at different airports may result from several factors. For instance, if there are large delays at a hub in the network, this may induce delays at other airports that have frequent connections to that hub. On the other

¹<https://www.bts.dot.gov/browse-statistical-products-and-data/bts-publications/airline-service-quality-performance-234-time>

hand, independently of flight connections, meteorological events such as severe storms or heavy snowfall can cause simultaneous delays at airports in the same geographical area.

In order to find groups of airports that have a risk of simultaneous large delays, we first run a k -medoids clustering (Kaufman and Rousseeuw, 2009), where similarities are defined in terms of the strength of extremal dependence. Clustering is frequently used in the extreme value literature to identify regions that are homogeneous in terms of dependence properties (e.g., Bernard et al., 2013; Saunders et al., 2020; Vignotto et al., 2021). We follow these approaches but use a different dissimilarity measure, namely the empirical version $\hat{\chi}_{ij}(p)$ of the extremal correlation (2.1) at probability level $p = 0.85$. The clusters and subsequent results are stable with respect to the exact probability level in a range of about $p \in (0.8, 0.95)$, guaranteeing both a sufficiently high threshold and enough exceedances.

The resulting clusters are shown in Figure 5, for an ad-hoc choice of six clusters. Even though no information on the locations of the airports is used, the resulting groups of airports exhibit strong geographical characteristics. This confirms our initial intuition on the importance of flight connectivity and meteorological influences. Figure S.2 shows that the extremal dependence of large delays is much stronger within the identified clusters than between different clusters. In the following we focus on the cluster in the South around Texas; similar analyses can be conducted for the other clusters. The existence of a (regular) flight connection between two airports defines a natural graph, which we denote by G_{flight} and show in the left panel of Figure 6.

6.3 Graphical modeling

The flight graph G_{flight} is based on domain knowledge and is certainly a good first candidate for statistical modeling. It is however not necessarily the best graph in terms of conditional independence properties. We therefore also consider extremal graph structures estimated in a data-driven way. In order to evaluate the fitted models out-of-sample, all rows containing missing values were removed, and the data set was split into a training set with 1974 observations (2005-01-01 to 2010-12-31) used for estimation, and a validation set with 2508 observations (2011-01-01 to 2020-12-31) used for selection of tuning parameters and model comparisons. As a base estimator, we use in the following the empirical extremal variogram $\hat{\Gamma}$ computed at probability threshold $p = 0.85$ on the training data set.

The first, sparse graph is obtained as the minimum spanning tree with weight matrix $\hat{\Gamma}$ as proposed in Engelke and Volgushev (2020); this tree is denoted by \hat{T} and is shown in the center of Figure 6. Alternatively, the empirical extremal correlation

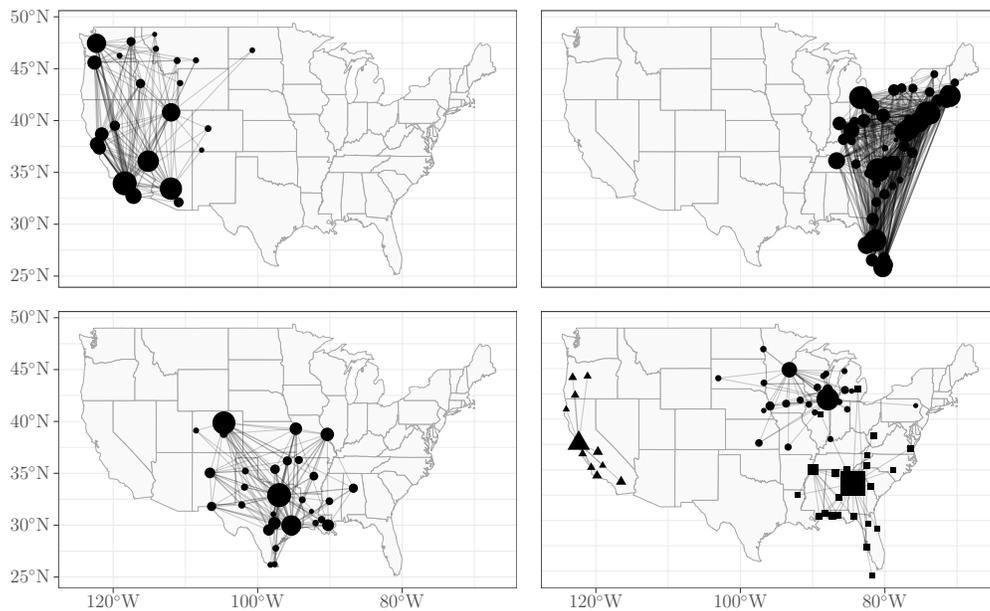


Figure 5: Clusters of airports in the contiguous U.S., using k -medoids clustering with empirical extremal correlation as dissimilarity measure. Two airports in a cluster are connected by edges if they are connected by (regular) flights. The smallest three clusters are shown together in the bottom-right panel, with airports in the same cluster represented by the same geometric shape. The size of an airport is proportional to the average number of daily flights at the location.

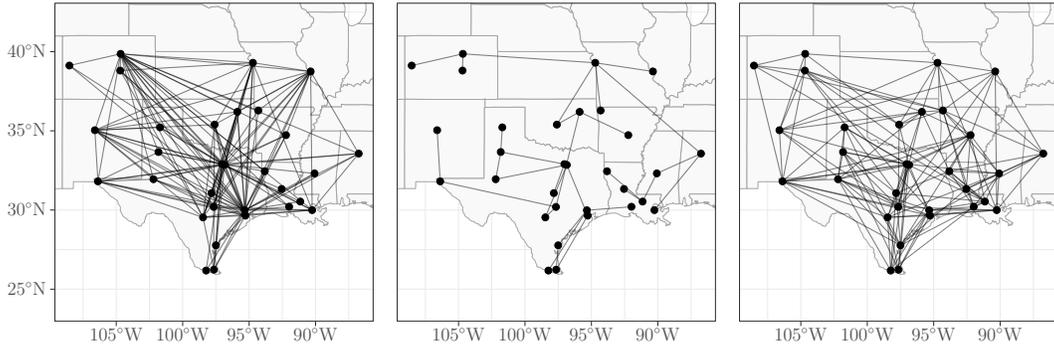


Figure 6: The flight graph G_{flight} (left), the estimated tree graph \hat{T} (center), and the graph estimated using **EGlearn**, G_{ρ} , for $\rho^* = 0.1$ (right).

$\hat{\chi}$ as weight matrix can also recover an underlying tree (Engelke and Volgushev, 2020; Hu et al., 2022).

As a second family of estimated extremal graph structures \hat{G}_{ρ} we apply the **EGlearn** algorithm (Engelke et al., 2022b) to the flights data set with different regularization parameters $\rho \geq 0$. The latter governs the amount of sparsity in the estimated graph, where larger ρ values correspond to sparser graphs. **EGlearn** produces a whole sequence of estimated graphical models, but without estimates of the corresponding parameter matrices.

We apply our methodology to fit a multivariate generalized Pareto distribution with Hüsler–Reiss distribution and extremal graphical structure given by different graphs G . As in the structure estimation, we focus on the empirical variogram $\hat{\Gamma}$ as base estimator. Applying the completion operator \mathfrak{C}_G to the restriction of $\hat{\Gamma}$ to a graph G as in (5.3), we obtain graph structured estimators $\hat{\Gamma}^{G_{\text{flight}}}$, $\hat{\Gamma}^{\hat{T}}$ and $\hat{\Gamma}^{\hat{G}_{\rho}}$ for the flight graph, the extremal tree and the ρ -regularized general graph, respectively. We note that previously, only parameter estimation on the tree was possible through the tree metric property (4.3), but not on the more general graphs. Our theory therefore produces a full statistical model that can be used for model assessment, simulation and interpretation.

A first benefit of our approach is that it complements structure learning approaches such as **EGlearn** by allowing for a data-driven selection of the optimal amount of sparsity. This can be done comparing the Hüsler–Reiss log-likelihood values of the different parameter matrices $\hat{\Gamma}^{\hat{G}_{\rho}}$, for instance using AIC or BIC if applied to the training data set. Instead, we directly compare the log-likelihood values on the validation set and plot them against the tuning parameter ρ in Figure 7. The likelihood is computed based on the Hüsler–Reiss Pareto density and since we

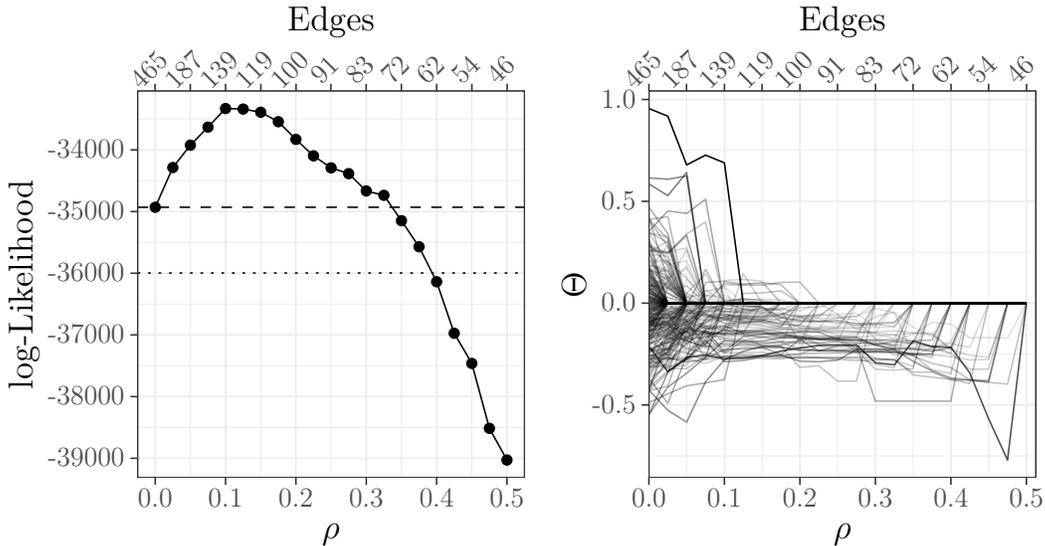


Figure 7: Left: Log-likelihood based on the validation data set for different regularization parameters ρ . Horizontal lines indicate the log-likelihoods of the complete graph (465 edges, dashed line) and the flight graph (151 edges, dotted line). The log-likelihood of the tree graph (30 edges) is -48681 and not shown. Right: Evolution of $\hat{\Theta}^{\hat{G}_\rho}$ entries. Only entries that vanish at $\rho = 0.5$ are shown. Both plots show the number of edges in the corresponding graph on top.

are mainly interested in the dependencies between different airports, the univariate marginals were normalized to the standard Pareto scale using the empirical distribution functions; Figure S.3 shows the univariate shape parameters before normalizing. The best fitting model on the independent validation data set comes from **EGlearn** at $\rho^* = 0.1$, and the corresponding graph with 139 edges is shown on the right-hand side of Figure 6. The flight graph performs significantly worse, and the tree seems to be too sparse. The Hüsler–Reiss precision matrix defined in Section 3.2 can be used to track how sparsity is induced in the **EGlearn** algorithm. Figure 7 shows the entries of the precision matrices $\hat{\Theta}^{\hat{G}_\rho}$ as a function of the tuning parameter ρ . Similar to a usual lasso, we observe that they tend to zero in a possibly non-monotone way.

In terms of model assessment, we can use the completed variogram matrices to evaluate the goodness of fit. Figure 8 compares the values of the empirical extremal variogram $\hat{\Gamma}$ to the variogram estimates implied by the fitted graphical model $\hat{\Gamma}^G$ for graph G being the flight graph, the extremal minimum spanning tree and the optimal **EGlearn** graph, respectively. Here we use the corresponding extremal correlations obtained as $\chi = 2 - 2\Phi(\sqrt{\Gamma}/2)$ for a Hüsler–Reiss distribution with variogram matrix

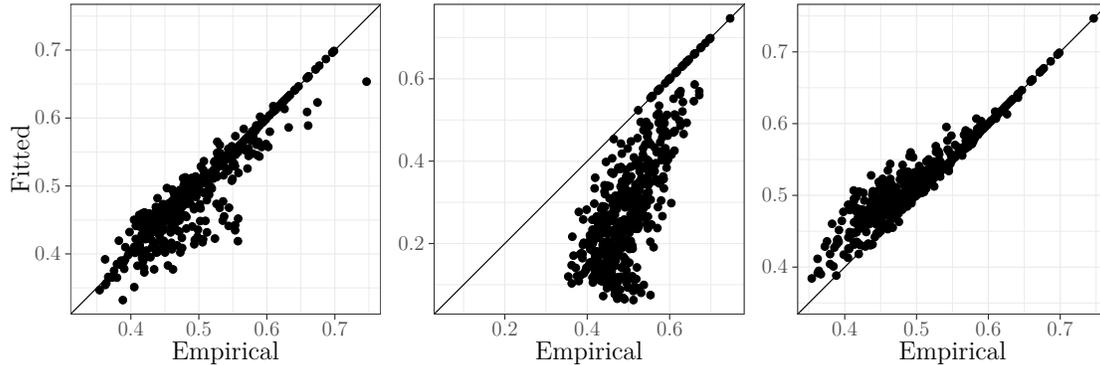


Figure 8: Extremal correlations based on the empirical and fitted extremal variogram for the flight graph G_{flight} (left), the estimated tree graph \hat{T} (center), and the graph estimated using `EGlearn`, G_ρ , for $\rho^* = 0.1$ (right).

Γ , where Φ is the standard normal distribution function and all functions are applied componentwise. The results confirm the likelihood considerations and, in particular, a tree model is clearly not flexible enough to model all extremal dependencies in this data set.

Acknowledgments

Financial support from the F.R.S-FNRS in the form of a PDR grant (T020321F, J. Segers) and the Swiss National Science Foundation (S. Engelke) is gratefully acknowledged.

References

- Asadi, P., Davison, A. C., and Engelke, S. (2015). Extremes on river networks. *Ann. Appl. Statist.*, 9(4):2023–2050.
- Asenova, S., Mazo, G., and Segers, J. (2021). 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.
- Asenova, S. and Segers, J. (2021). Extremes of Markov random fields on block graphs. Available from <https://arxiv.org/abs/2112.04847>.

- Bakonyi, M. and Johnson, C. R. (1995). The euclidian distance matrix completion problem. *SIAM J. Matrix Anal. Appl.*, 16(2):646–654.
- Bakonyi, M. and Woerdeman, H. J. (2011). *Matrix Completions, Moments, and Sums of Hermitian Squares*. Princeton University Press.
- Baz, J., Díaz, I., Montes, S., and Pérez-Fernández, R. (2022). Some results on the gaussian markov random field construction problem based on the use of invariant subgraphs. *TEST*.
- Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: spatial dependencies among heavy rainfall in france. *J Climate*, 26(20):7929–7937.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.
- Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.
- de Fondeville, R. and Davison, A. C. (2018). High-dimensional peaks-over-threshold inference. *Biometrika*, 105:575–592.
- de Haan, L. and Resnick, S. (1977). Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete*, 40:317–337.
- Dombry, C., Engelke, S., and Oesting, M. (2016). Exact simulation of max-stable processes. *Biometrika*, 103:303–317.
- Dombry, C., Eyi-Minko, F., and Ribatet, M. (2013). Conditional simulation of max-stable processes. *Biometrika*, 100(1):111–124.
- Einmahl, J. H. J., Kiriliouk, A., Krajina, A., and Segers, J. (2016). An M-estimator of spatial tail dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol*, 78:275–298.
- Einmahl, J. H. J., Kiriliouk, A., and Segers, J. (2018). A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, 21(2):205–233.
- Einmahl, J. H. J., Krajina, A., and Segers, J. (2012). An M-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.*, 40:1764–1793.
- Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol*, 82(4):871–932.

- Engelke, S., Hitz, A. S., Gnecco, N., and Hentschel, M. (2022a). *graphicalExtremes: Statistical Methodology for Graphical Extreme Value Models*. Available from <https://github.com/sebastian-engelke/graphicalExtremes>, R package version 0.1.0.9000.
- Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annu. Rev. Stat. Appl.*, 8:241–270.
- Engelke, S., Lalancette, M., and Volgushev, S. (2022b). Learning extremal graphical structures in high dimensions. Available from <https://arxiv.org/abs/2111.00840>.
- Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2015). Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 77(1):239–265.
- Engelke, S. and Volgushev, S. (2020). Structure learning for extremal tree models. Available from <https://arxiv.org/abs/2012.06179>.
- Fang, H.-r. and O’Leary, D. P. (2012). Euclidean distance matrix completion problems. *Optimization Methods and Software*, 27(4-5):695–717.
- Farris, J. S., Kluge, A. G., and Eckardt, M. J. (1970). A numerical approach to phylogenetic systematics. *Syst Zool.*, 19(2):172–189.
- Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24:2693–2720.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, third edition.
- Gower, J. (1982). Euclidean distance geometry. *Math. Scientist*, 7:1–14.
- Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(3):497–546.
- Ho, Z. W. O. and Dombry, C. (2019). Simple models for multivariate regular variation and the Hüsler–Reiss Pareto distribution. *J. Multivariate Anal.*, 173:525–550.
- Hu, S., Peng, Z., and Segers, J. (2022). Modelling multivariate extreme value distributions via markov trees. Available from <https://arxiv.org/abs/2208.02627>.

- Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: Between independence and complete dependence. *Statist. Prob. Letters*, 7(4):283–286.
- Kabluchko, Z., Schlather, M., and de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Probab.*, 37(5):2042 – 2065.
- Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.
- Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61:123–135.
- Knill, O. (2013). Cauchy–Binet for pseudo-determinants. *Linear Algebra Appl.*, 459.
- Kulpa, W. (1998). Poincaré and domain invariance theorem. *Acta Univ. Carolin. Math. Phys.*, 39(1-2):127–136.
- Lalancette, M., Engelke, S., and Volgushev, S. (2021). Rank-based estimation under asymptotic dependence and independence, with applications to spatial extremes. *Ann. Statist.*, 49(5):2552 – 2576.
- Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford statistical science series*. Clarendon Press, Oxford.
- Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187.
- Lee, D. and Joe, H. (2018). Multivariate extreme value copulas with factor and tree dependence structures. *Extremes*, 21:147–176.
- Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.*, 105:263–277.
- Papastathopoulos, I., Strokorb, K., Tawn, J. A., and Butler, A. (2017). Extreme events of Markov chains. *Adv. in Appl. Probab.*, 49(1):134–161.
- Penrose, R. (1955). A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.*, 51(3):406–413.
- Rakočević, V. (1997). On continuity of the Moore-Penrose and Drazin inverses. *Mat. Vesnik*, 49.

- Rao, C. R. (2002). *Linear statistical inference and its applications*. Wiley series in probability and statistics. Wiley, New York, 2. ed., paperback ed edition.
- Reich, B. J., Shaby, B. A., and Cooley, D. (2014). A hierarchical model for serially-dependent extremes: A study of heat waves in the western US. *Journal of Agricultural, Biological, and Environmental Statistics*, 19:119–135.
- Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *J. Multivariate Anal.*, 165:117–131.
- Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12:917–930.
- Röttger, F., Engelke, S., and Zwiernik, P. (2021). Total positivity in multivariate extremes. Available from <https://arxiv.org/abs/2112.14727>.
- Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, volume 104 of *Monographs on statistics and applied probability*. Chapman & Hall/CRC, Boca Raton, Fla.
- Saunders, K. R., Stephenson, A. G., and Karoly, D. J. (2020). A regionalisation approach for rainfall based on extremal dependence. *Extremes*.
- Segers, J. (2020). One- versus multi-component regular variation and extremes of Markov trees. *Adv. in Appl. Probab.*, 52:855–878.
- Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.*, 14(1):138–150.
- Uhler, C. (2017). Gaussian graphical models: An algebraic and geometric perspective. Available from <https://arxiv.org/abs/1707.04345>.
- Vignotto, E., Engelke, S., and Zscheischler, J. (2021). Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain and Ireland. *Weather Clim. Extremes*, 32:100318.
- Wadsworth, J. and Tawn, J. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika*, 101(1):1–15.
- Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

SUPPLEMENTARY MATERIAL

S.1 Moore–Penrose inverse and pseudo-determinant

S.1.1 Moore–Penrose inverse

Definition S.1.1. *As shown for example in Penrose (1955, Theorem 1), for any matrix $A \in \mathbb{R}^{n \times m}$ there exists a unique matrix $B \in \mathbb{R}^{m \times n}$ satisfying the equations*

$$\begin{aligned}ABA &= A, & (AB)^\top &= AB, \\BAB &= B, & (BA)^\top &= BA.\end{aligned}$$

This solution is called the Moore–Penrose inverse or simply “pseudo-inverse” of A and is denoted by $A^+ := B$.

Remark S.1.2. The pseudo-inverse is defined in a similar way for matrices with complex-valued entries, using the conjugate transpose in place of the transpose.

Lemma S.1.3. *The pseudo-inverse has the following properties.*

- (1) $(A^+)^+ = A$
- (2) $(AA^\top)^+ = (A^+)^\top A^+$ (note that in general $(AB)^+ \neq B^+ A^+$).
- (3) AA^+ is the orthogonal projection onto the image of A .
- (4) Using the singular value decomposition $A = U\Sigma V^\top$, the pseudo-inverse can be computed as $A^+ = V\Sigma^+ U^\top$, with

$$\Sigma^+ = \text{diag}(\Sigma_{11}^{-1}, \dots, \Sigma_{rr}^{-1}, 0, \dots, 0), \quad r = \text{rank}(A).$$

Proof. The first statement follows from the interchangeability of A and B in Definition S.1.1. The second statement can be proven by substituting the left- and right-hand side in the defining equations from Definition S.1.1. The last two statements are from Golub and Van Loan (1996, Section 5.5.4). \square

Lemma S.1.4. *Let A and B be two symmetric matrices of equal size and let P_A denote the orthogonal projection matrix onto the image of A . Then the following two conditions are sufficient and necessary for $A = B^+$:*

- $\text{Im } B \subseteq \text{Im } A$ (or equivalently $\ker A \subseteq \ker B$);
- $AB = P_A$.

Proof. In the first bullet point, the equivalence of $\text{Im } B \subseteq \text{Im } A$ and $\ker A \subseteq \ker B$ follows from the identity $(\ker M)^\perp = \text{Im } M$ for symmetric matrices M .

For $A = B^+$, the two conditions follow directly from Lemma S.1.3; note that by the lemma and by symmetry of A and B , we have $AB = (AB)^\top = B^\top A^\top = BA$.

Conversely, using the fact that the projection matrix P_A is symmetric, the defining equations from Definition S.1.1 follow from the two conditions as follows:

$$\begin{aligned} AB &= P_A = (P_A)^\top = (AB)^\top, \\ BA &= B^\top A^\top = (AB)^\top = AB = A^\top B^\top = (BA)^\top, \\ ABA &= P_A A = A, \\ BAB &= (AB)B = P_A B = B, \end{aligned}$$

where we used $\text{Im } B \subseteq \text{Im } A$ in the last step. □

S.1.2 Pseudo-determinant

Definition S.1.5 and Lemma S.1.6 are from Knill (2013, Section 2) and only adapted in scope and notation for their use here.

Definition S.1.5. *Let A be a square matrix with eigenvalues $\{\lambda_i\}$. Then its pseudo-determinant, denoted by $|A|_+$, is defined as the product of its non-zero eigenvalues:*

$$|A|_+ = \prod_{\lambda_i \neq 0} \lambda_i.$$

If all eigenvalues are zero, $|A|_+ = 1$.

Lemma S.1.6. *Let $A, B \in \mathbb{R}^{d \times d}$.*

- (1) *If A is similar to B , then $|A|_+ = |B|_+$.*
- (2) *If A is invertible then $|A|_+ = |A|$.*
- (3) $|A^\top|_+ = |A|_+$.
- (4) *For a normal matrix A , it holds that $|A^+|_+ = 1/|A|_+$.*
- (5) $|A^\top B|_+ = |AB^\top|_+ = |BA^\top|_+$.

(6) $|A|_+ \neq 0$.

(7) If A is block diagonal, i.e., if $A = \text{Diag}(A_1, \dots, A_k)$, then $|A|_+ = \prod_i |A_i|_+$.

Lemma S.1.7. Let V be a linear subspace of \mathbb{R}^d and put

$$\mathcal{Q}_V = \{A \in \mathbb{R}^{d \times d} : A = A^\top, \ker A = V\}.$$

Then for any $A_1, A_2 \in \mathcal{Q}_V$,

$$|A_1 A_2|_+ = |A_1|_+ \cdot |A_2|_+.$$

Furthermore, for a sequence $A_i \in \mathcal{Q}_V$, $i \in \mathbb{N}$, with $\lim_{i \rightarrow \infty} A_i = A_0 \in \mathcal{Q}_V$,

$$\lim_{i \rightarrow \infty} |A_i|_+ = |A_0|_+.$$

Proof. Let $k = \dim V$, $l = d - k$, and let $\{b_1, \dots, b_d\}$ be a basis of \mathbb{R}^d with $V = \text{span}(b_1, \dots, b_k)$. Let M denote the corresponding basis change matrix and $B_i = M^{-1} A_i M$ for $i \geq 0$. By construction, B_i is of the form

$$\begin{pmatrix} \mathbf{0}_{k \times k} & \mathbf{0}_{k \times l} \\ \mathbf{0}_{l \times k} & B'_i \end{pmatrix}$$

with $B'_i \in \mathbb{R}^{l \times l}$ invertible and $\mathbf{0}_{a \times b} \in \mathbb{R}^{a \times b}$ equal to zero in all entries. Using the properties from Lemma S.1.6 it follows that

$$|A_i|_+ = |B_i|_+ = |\mathbf{0}_{k \times k}|_+ \cdot |B'_i|_+ = |B'_i|_+.$$

Since M is chosen independently of i ,

$$\begin{aligned} |A_1|_+ \cdot |A_2|_+ &= |B'_1|_+ \cdot |B'_2|_+ \\ &= |B'_1 B'_2|_+ \\ &= |B_1 B_2|_+ \\ &= |M B_1 M^{-1} M B_2 M^{-1}|_+ \\ &= |A_1 A_2|_+. \end{aligned}$$

Furthermore, continuity of matrix multiplication implies that $B_0 = \lim_{i \rightarrow \infty} B_i$ and hence $\lim_{i \rightarrow \infty} B'_i = B'_0$. Using the continuity of the regular determinant, it follows that

$$\lim_{i \rightarrow \infty} |A_i|_+ = \lim_{i \rightarrow \infty} |B'_i|_+ = |B'_0|_+ = |A_0|_+. \quad \square$$

S.2 Graph theory

All graphs considered in this paper are undirected, simple graphs without loops. A graph $G = (V, E)$ is defined by a set of vertices $V = \{1, \dots, d\}$ and a set of edges $E \subseteq V \times V$. Since we only consider undirected graphs without loops, the set of all possible edges is $\mathcal{E}(V) := (V \times V) \setminus \{(i, i) : i \in V\}$, and it must always hold that $(i, j) \in E \Rightarrow (j, i) \in E$. The set of loops from each node to itself is denoted as $D_V = \{(i, i) : i \in V\}$. These loops are no edges in the sense defined above, but can be useful when identifying edges and subgraphs with matrix entries and submatrices, respectively. For a set of edges $E \subseteq \mathcal{E}(V)$, the inclusion of the loops from each node to itself is denoted as $\overline{E} = E \cup D_V$.

A graph is called complete if $E = \mathcal{E}(V)$. A subgraph $G' = (V', E')$ of G is a graph consisting of a subset of vertices $V' \subseteq V$ and a subset of E such that all endpoints lie in V' , i.e., $E' \subseteq E \cap \mathcal{E}(V')$. If E' is maximal (i.e., the latter set inclusion is an equality), G' is called the subgraph induced by V' . A subset of vertices $C \subseteq V$ is called complete if its induced subgraph is a complete graph. A subset of vertices is called a clique if it is complete and not a strict subset of another complete subset.

The neighborhood of a vertex i is defined as $\delta(i) = \{j \in V : (i, j) \in E\}$. The neighborhood of a vertex including the vertex itself is denoted as $\bar{\delta}(i) := \delta(i) \cup \{i\}$. A path of length m between vertices i and j is a sequence of $m + 1$ distinct vertices p_0, p_1, \dots, p_m such that $p_0 = i$, $p_m = j$, and $(p_{i-1}, p_i) \in E$ for all $i = 1, \dots, m$. If there exists a path between two vertices, they are said to be connected. A graph is connected if any two of its vertices are connected. A cycle of length m is a sequence of m distinct vertices p_1, \dots, p_m such that $(p_m, p_1) \in E$ and $(p_{i-1}, p_i) \in E$ for all $i = 2, \dots, m$. A chord is an edge between two vertices of a cycle that is not itself part of the cycle.

Definition S.2.1 (Decomposable Graph). *A decomposable graph is a graph in which all cycles of four or more vertices have a chord.*

A useful property of decomposable graphs is the following running intersection property (see e.g., Lauritzen, 1996).

Lemma S.2.2 (Running intersection property). *For a decomposable graph G , the set of cliques $\mathcal{C} = \{C_1, \dots, C_N\}$ can be ordered such that the running intersection property is fulfilled, that is, for all $i = 2, \dots, N$ there exists $k(i) \in \{1, \dots, i - 1\}$ such that*

$$D_i := C_i \cap (C_1 \cup \dots \cup C_{i-1}) \subseteq C_{k(i)}.$$

The multiset $\mathcal{D} = \{D_2, \dots, D_N\}$ is independent of the chosen ordering of \mathcal{C} and its elements are called separators. For connected graphs, all separators are non-empty.

Definition S.2.3 (Block graph). *A block graph is a decomposable graph in which all non-empty separators consist of single vertices:*

$$|D| \in \{0, 1\} \quad \forall D \in \mathcal{D}.$$

Definition S.2.4 (Tree Graph). *A tree graph or a tree is a connected graph that does not contain any cycle.*

Remark S.2.5. For $d \geq 2$, the set of trees is identical to the set of connected block graphs in which all cliques consist of exactly two vertices.

Definition S.2.6 (Graph Laplacian). *For an undirected graph $G = (V, E)$ the graph Laplacian matrix $L \in \mathbb{R}^{d \times d}$ is defined by*

$$L_{ij} = \begin{cases} \deg(i) & i = j, \\ -1 & (i, j) \in E, \\ 0 & \text{otherwise,} \end{cases}$$

where the degree $\deg(i)$ of a vertex i is defined as $|\delta(i)|$.

S.3 The Hüsler–Reiss exponent measure density

Using the general representation in (3.10), we recover different forms of the Hüsler–Reiss density from the literature. Canonical unit vectors are a natural choice for the vector v . Setting $v = e_k$ for some $k \in V$, yields the density representation in Definition 3.1. The corresponding stochastic representation (3.11) results in

$$Y^{(k)} = W^{(k)} + R \cdot \mathbf{1},$$

where $Y^{(k)}$ is the random vector used in Section 2.2 and $W^{(k)} = W_{e_k}$ is called the k th extremal function (Dombry et al., 2013). This stochastic representation coincides with Expression (28) of Engelke and Hitz (2020). The random vector $W^{(k)}$ is used for various purposes including statistical inference (Engelke et al., 2015), exact simulation (Dombry et al., 2016), and the definition of extremal positive dependence (Röttger et al., 2021).

Another characterization of the Hüsler–Reiss density is given in Wadsworth and Tawn (2014), parametrized by a positive definite matrix Σ with a constant diagonal, as

$$\lambda(y; \Sigma) = c_\Sigma \cdot \exp\left(-\frac{1}{2}y^\top A y - y^\top q (\mathbf{1}^\top q)^{-1}\right), \quad y \in \mathbb{R}^d,$$

with $q = \Sigma^{-1}\mathbf{1}$ and $A = \Sigma^{-1} - qq^\top(\mathbf{1}^\top q)^{-1}$. Notably, it can be verified (e.g., using Lemma S.1.4) that $A = (\Pi\Sigma\Pi)^\dagger = \theta \circ \gamma(\Sigma)$, i.e., A is equal to our precision matrix Θ for the positive semi-definite covariance matrix $\Pi\Sigma\Pi \in \mathcal{P}_d^1$.

Likewise, a slightly more general definition of the Hüsler–Reiss exponent measure density is given by Ho and Dombry (2019, Definition 3.1) as

$$\lambda(y; a, Q, l) = c_{a,Q,l} \cdot \exp\left(-\frac{1}{2}y^\top Q y + l^\top y\right),$$

with some $a \in (0, \infty)^d$, $Q \in \mathcal{P}_d^1$, and $l^\top \mathbf{1} < 0$, defined for $y \in \mathbb{R}^d$, and the corresponding multivariate generalized Pareto distribution supported on $\{y \in \mathbb{R}^d : \exp(y) \not\leq a\}$. Using this definition, the authors show an exponential family property of the Hüsler–Reiss Pareto model. Restricting the parameters to $a = \mathbf{1}$ and $l = -(r_\Theta + \mathbf{e}_d)$ yields an equivalent parameter space to the one spanned by the set of conditionally negative definite matrices in Definition 3.1. The matrix Q is identical to the precision matrix Θ , possibly allowing for an easier interpretation of the exponential family results in the framework of extremal graphical models. Kiriliouk et al. (2018, Section 7.2) give a definition that is equivalent to the one above, restricted to $a = \mathbf{1}$ and $l^\top \mathbf{1} = -1$.

To allow for a comparison of the parametrizations in (3.9) and (3.10) with the general definition in Ho and Dombry (2019), we suggest the following parametrization of their (general) Hüsler–Reiss exponent measure density.

Definition S.3.1. *Let $\Theta \in \mathcal{P}_d^1$, $\alpha > 0$, and $r \in \{\mathbf{1}\}^\perp$. For $y \in \mathbb{R}^d$ define the general Hüsler–Reiss exponent measure density as*

$$\lambda(y; \Theta, \alpha, r) \propto \exp\left(-\frac{1}{2}y^\top \Theta y - \alpha y^\top \mathbf{e}_d + y^\top r\right),$$

or equivalently

$$\lambda(y; \Theta, \alpha, r) \propto \exp(-\alpha y^\top v) \cdot \exp\left(-\frac{1}{2}\|y - \mu_v\|_\Theta^2\right),$$

with $v \in \mathbb{R}^d$, satisfying $v^\top \mathbf{1} = 1$, and $\mu_v = \Theta^+(r + \alpha v)$.

S.4 Details on matrix completion

S.4.1 Matrix completion as likelihood optimization

For a Hüsler–Reiss distribution Y with parameter matrix Γ , the random vectors $Y^{(k)}$, $k \in V$, defined in Section 2.2 satisfy

$$(Y_i^{(k)} - Y_k^{(k)})_{i \neq k} \sim \mathcal{N}\left(-\frac{1}{2} \text{diag}(\Sigma^{(k)}), \Sigma^{(k)}\right); \quad (\text{S.4.1})$$

this follows from representation (3.11) and Corollary 3.7 with $v = e_k$ (see also Engelke et al., 2015). For an independent sample of size n from Y , one can obtain samples from this $(d - 1)$ -dimensional normal distribution by selecting only data with $Y_k > 1$ and following (S.4.1). We denote the corresponding empirical covariance matrix by $\widehat{\Sigma}^{(k)}$, augmented by a k th row and column of zeros to make a $d \times d$ matrix. Ignoring the information on the parameter matrix Γ in the mean vector, the (surrogate) log-likelihood of this model can be written in terms of our Hüsler–Reiss precision matrix Θ as

$$L(\Theta; \widehat{\Sigma}^{(k)}) \propto \log |\Theta|_+ - \text{tr}(\widehat{\Sigma}^{(k)}\Theta), \quad (\text{S.4.2})$$

where $|\cdot|_+$ is the pseudo-determinant. Setting $\widehat{\Gamma}^{(k)} = \gamma(\widehat{\Sigma}^{(k)})$ gives a nonparametric estimator of Γ . By Proposition 3.3 it holds that $\widehat{\Sigma}^{(k)} = \Pi(-\frac{1}{2}\widehat{\Gamma}^{(k)})\Pi$ and therefore the cyclic permutation property of the trace operator together with the fact that $\Theta\Pi = \Pi\Theta = \Theta$ for any $\Theta \in \mathcal{P}_d^1$ shows that the right-hand side of (S.4.2) is equal to

$$\log |\Theta|_+ + \frac{1}{2}\text{tr}(\overline{\Gamma}\Theta),$$

with $\overline{\Gamma} = \widehat{\Gamma}^{(k)}$. In order to use all data in the sample, we can consider this likelihood with $\overline{\Gamma}$ equal to a combined version of the variogram estimators over all k , defined as $\widehat{\Gamma} := d^{-1} \sum_{k=1}^d \widehat{\Gamma}^{(k)}$ and called the empirical variogram (Engelke and Volgushev, 2020).

A natural way to obtain a graphical model is therefore to maximize this surrogate Hüsler–Reiss likelihood over $\Theta \in \mathcal{P}_d^1$ under the constraint of a graph-structured precision matrix. As shown in Proposition 5.1, solving this optimization problem corresponds to our completion operator \mathfrak{C}_G from Definition 4.6.

S.4.2 Estimation strategy for sparse graphs

Let \mathcal{C} be the collection of all cliques $C \subseteq V$ of the graph $G = (V, E)$ and suppose that for every $C \in \mathcal{C}$, $\widehat{\Gamma}_C$ is a consistent estimator of the entries Γ_{ij} for $i, j \in C$. Note that $\widehat{\Gamma}_C$ can be computed using information only from the components in C , which reduces the computational cost drastically if the cliques are small, especially if censoring is used. Since the cliques are overlapping on the separator sets, we need to combine different estimators to obtain a partial variogram matrix $\widehat{\Gamma} \in \mathring{\mathbb{R}}^{d \times d}$ on the whole graph. We do this by averaging the estimators on the intersections. Let $\mathcal{C}_{(i,j)} = \{C \in \mathcal{C} : i, j \in C\}$ denote the set of all cliques containing the edge $(i, j) \in E$

and put

$$\hat{\Gamma}_{ij} = \begin{cases} \text{“?”} & (i, j) \notin \bar{E}, \\ \frac{1}{|\mathcal{C}_{(i,j)}|} \sum_{C \in \mathcal{C}_{(i,j)}} (\hat{\Gamma}_C)_{ij} & (i, j) \in \bar{E}. \end{cases}$$

This partial variogram estimator is clearly consistent for $\Gamma|_G$ as in (5.2), since it is an average of consistent estimators. The completion $\hat{\Gamma}^G = \mathfrak{C}_G(\hat{\Gamma})$ from Corollary 5.4 is then a consistent estimator of Γ with correct graph structure G . For fixed sample size n , the estimator $\hat{\Gamma}$ is not guaranteed to be a valid (partial) variogram matrix; this is the price to pay for the more efficient estimation using only information within each clique, and by Theorem 5.3, the probability of $\hat{\Gamma}$ being invalid converges to zero for $n \rightarrow \infty$.

A natural question is whether it is possible to replace the arithmetic mean by another function that guarantees a valid partial variogram. The following example shows that this is not possible, as long as entries that belong to only a single clique are not altered, as well.

Example S.4.1. Let

$$\hat{\Gamma} = \begin{pmatrix} 0 & 1 & 16 & ? \\ 1 & 0 & x & 1 \\ 16 & x & 0 & 1 \\ ? & 1 & 1 & 0 \end{pmatrix}.$$

Then there are $x_1, x_2 \in \mathbb{R}$ such that with $x = x_1$ the principal submatrix $\hat{\Gamma}|_{\{1,2,3\}}$ is conditionally negative definite, and with $x = x_2$ the principal submatrix $\hat{\Gamma}|_{\{2,3,4\}}$ is conditionally negative definite, but there exists no x such that both submatrices are conditionally negative definite at the same time.

Proof. Gower (1982) shows that Γ can be interpreted as a Euclidean distance matrix with $\Gamma_{ij} = d_{ij}^2$, where d_{ij} , for $i, j \in \{1, \dots, d\}$, is the pairwise distance between the points generating Γ (see also Section S.5.4.3). The triangle inequality then requires $9 \leq x_1 \leq 25$ and $0 \leq x_2 \leq 4$, which can be satisfied for each submatrix but not simultaneously for a single value of x . \square

S.5 Proofs

S.5.1 Proof of Proposition 3.3

Before proving the Proposition we establish some auxiliary results. In the following, for a given $\Gamma \in \mathcal{D}_d$ we write $\Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi$. We assume that the matrix $S \in \mathbb{R}^{d \times d}$ satisfies $\Pi S \Pi = \Sigma$, the matrix S not necessarily being symmetric. Lemma S.5.8 shows that this is in fact a slightly weaker assumption than $\gamma(S) = \Gamma$, which is used in the Proposition. Furthermore, for any $t \in \mathbb{R}$ we introduce the abbreviation

$$\Sigma^{[t]} = t\mathbf{1}\mathbf{1}^\top + S.$$

Lemma S.5.1. *The matrix Σ is symmetric and positive semi-definite, and its kernel is $\text{span}(\{\mathbf{1}\})$.*

Proof. Since $\Pi = I_d - d^{-1}\mathbf{1}\mathbf{1}^\top$ is the projection matrix onto the subspace orthogonal to $\mathbf{1}$, it follows that $\Sigma\mathbf{1} = \Pi(-\frac{1}{2}\Gamma)(\Pi\mathbf{1}) = \mathbf{0}$. Hence, $\text{span}(\{\mathbf{1}\}) \subseteq \ker \Sigma$. Next, consider any $v \notin \text{span}(\{\mathbf{1}\})$, and let $w := \Pi v$. This w satisfies $w \neq \mathbf{0}$ and $w \perp \mathbf{1}$, hence

$$\begin{aligned} v^\top \Sigma v &= v^\top (\Pi(-\frac{1}{2}\Gamma)\Pi)v \\ &= -\frac{1}{2}w^\top \Gamma w \\ &> 0, \end{aligned}$$

because of the conditional negative definiteness in the definition of a variogram matrix in (3.1). Symmetry follows from the symmetry of Γ and Π . \square

Lemma S.5.2. *There exists $t_0 \in \mathbb{R}$ such that $\Sigma^{[t]}$ is invertible for all $t \neq t_0$. Specifically, $t_0 = -(\mathbf{1}^\top S^{-1}\mathbf{1})^{-1}$ if S is invertible and $t_0 = 0$ otherwise.*

Proof. To prove the invertibility of $\Sigma^{[t]}$ for $t \neq t_0$ consider the following two cases.

Case 1: S is invertible. Let $v_0 = S^{-1}\mathbf{1}$. Then $\mathbf{1}^\top v_0 \neq 0$, because otherwise, if $\mathbf{1}^\top v_0 = 0$, then v_0 would be orthogonal to the kernel of Σ , and we would have $\Pi v_0 = v_0$, and therefore

$$\mathbf{0} \neq \Sigma v_0 = \Pi S \Pi v_0 = \Pi S v_0 = \Pi \mathbf{1} = \mathbf{0},$$

a contradiction. Next, let $t_0 = -(\mathbf{1}^\top v_0)^{-1}$. Consider $t \neq t_0$ and suppose that $\Sigma^{[t]}$ is

singular, i.e., there exists $v \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ such that $\mathbf{0}^\top = v^\top(t\mathbf{1}\mathbf{1}^\top + S)$. Then

$$\begin{aligned} 0 &= v^\top(t\mathbf{1}\mathbf{1}^\top + S)v_0 \\ &= tv^\top\mathbf{1}\mathbf{1}^\top v_0 + v^\top S v_0 \\ &= t(v^\top\mathbf{1})(\mathbf{1}^\top v_0) + (v^\top\mathbf{1}) \\ &= (v^\top\mathbf{1})(-t \cdot t_0^{-1} + 1). \end{aligned}$$

For $t \neq t_0$, the second factor cannot be zero, so it must hold that $\mathbf{1}^\top v = 0$. However, this leads to the contradiction

$$\mathbf{0} = (t\mathbf{1}\mathbf{1}^\top + S)v = Sv \neq \mathbf{0}.$$

Furthermore, it holds that

$$\Sigma^{[t_0]}v_0 = \left(S - \frac{\mathbf{1}\mathbf{1}^\top}{\mathbf{1}^\top S^{-1}\mathbf{1}}\right)S^{-1}\mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}.$$

Case 2: S is singular. Then S must be of rank $d - 1$, since the rank of $\Sigma = \Pi S \Pi$ is $d - 1$ by Lemma S.5.1. Let $v_0 \neq \mathbf{0}$ be such that $\ker S = \text{span}(\{v_0\})$. Then, $v_0 \notin \{\mathbf{1}\}^\perp$, because otherwise

$$\mathbf{0} \neq \Sigma v_0 = \Pi S \Pi v_0 = \Pi S v_0 = \mathbf{0}.$$

Furthermore, $\mathbf{1} \notin \text{Im } S$, because otherwise there must be a vector $u \notin \text{span}(\{v_0\})$, such that $Su = \mathbf{1}$. Then, $(\Pi S)v_0 = (\Pi S)u = \mathbf{0}$, implying that the rank of $\Pi S \Pi$ is at most $d - 2$, which is a contradiction.

Setting $t_0 := 0$, and using $\text{Im } S = S\mathbb{R}^d = S(\{\mathbf{1}\}^\perp \oplus v_0\mathbb{R}) = S\{\mathbf{1}\}^\perp$, the image of $\Sigma^{[t]}$ for $t \neq t_0$ can then be checked to be

$$\begin{aligned} \text{Im } \Sigma^{[t]} &= \Sigma^{[t]}\mathbb{R}^d \\ &= \Sigma^{[t]}(\{\mathbf{1}\}^\perp \oplus v_0\mathbb{R}) \\ &= (\Sigma^{[t]}\{\mathbf{1}\}^\perp) \oplus (\Sigma^{[t]}v_0\mathbb{R}) \\ &= (S\{\mathbf{1}\}^\perp) \oplus (t\mathbf{1}\mathbf{1}^\top v_0\mathbb{R}) \\ &= \text{Im } S \oplus \mathbf{1}\mathbb{R} \\ &= \mathbb{R}^d, \end{aligned}$$

implying that $\Sigma^{[t]}$ is invertible. Again, it holds that $\Sigma^{[t_0]}v_0 = Sv_0 = \mathbf{0}$. □

Lemma S.5.3. $\Sigma^{[t]}$ is positive definite (not necessarily symmetric) for all $t > t_0$.

Proof. A matrix $A \in \mathbb{R}^{d \times d}$ is positive (semi-)definite if and only if its symmetric part $\check{A} = \frac{1}{2}(A + A^\top)$ is so; indeed, $x^\top A x = x^\top A^\top x = x^\top \check{A} x$ for any $x \in \mathbb{R}^d$. The symmetric part of $\Sigma^{[t]}$ is $\check{\Sigma}^{[t]} = \frac{1}{2}(\Sigma^{[t]} + (\Sigma^{[t]})^\top) = t\mathbf{1}\mathbf{1}^\top + \check{S}$ where $\check{S} = \frac{1}{2}(S + S^\top)$ is the symmetric part of S . Further, since $\Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi$ is symmetric, the matrix $S \in \mathbb{R}^{d \times d}$ satisfies $\Pi S \Pi = \Sigma$ if and only if $\Pi S^\top \Pi = \Sigma$ and thus if and only if $\Pi \check{S} \Pi = \Sigma$. Hence, to show that $\Sigma^{[t]}$ is positive definite for $t > t_0$, we can without loss of generality assume S is symmetric (or more precisely, replace S by its symmetric part).

So assume S is symmetric and let v_0 be the vector from the proof of Lemma S.5.2. First, we show that $\Sigma^{[t_0]}$ is positive semi-definite; to do so, it is sufficient to show that its non-zero eigenvalues are positive. To this end, recall that $\Sigma^{[t_0]}v_0 = \mathbf{0}$ and consider an eigenvector $v_1 \notin \text{span}(\{v_0\})$ of $\Sigma^{[t_0]}$ with eigenvalue $\alpha_1 \neq 0$. The vector

$$u = (\mathbf{1}^\top v_0)v_1 - (\mathbf{1}^\top v_1)v_0$$

satisfies $u \perp \mathbf{1}$ and thus $\Pi u = u$. Moreover, since $\mathbf{1}^\top v_0 \neq 0$, it follows that $u \neq \mathbf{0}$. Hence, since Γ is conditionally negative definite and since v_0 and v_1 are orthogonal (as eigenvectors of the symmetric matrix $\Sigma^{[t_0]}$ associated to distinct eigenvalues), we have

$$\begin{aligned} 0 < u^\top \left(-\frac{1}{2}\Gamma\right) u &= u^\top \Pi \left(-\frac{1}{2}\Gamma\right) \Pi u = u^\top S u = u^\top \Sigma^{[t_0]} u \\ &= (\mathbf{1}^\top v_0)^2 \|v_1\|^2 \alpha_1 \end{aligned}$$

implying $\alpha_1 > 0$, as required. A similar argument shows that $\Sigma^{[t_0]}$ has rank $d - 1$: otherwise, we could find an eigenvector $v_1 \notin \text{span}(\{v_0\})$ of $\Sigma^{[t_0]}$ orthogonal to v_0 with eigenvalue 0 as well, and this would lead to a contradiction by the same calculation as above. For $t > t_0$, the matrix $\Sigma^{[t]}$ is invertible by Lemma S.5.2, and since it is the sum of the two positive semi-definite matrices $\Sigma^{[t_0]}$ and $(t - t_0)\mathbf{1}\mathbf{1}^\top$, it is also positive semi-definite and hence positive definite. \square

Lemma S.5.4. The limit $\lim_{t \rightarrow \infty} (\Sigma^{[t]})^{-1}$ exists, is symmetric, and its kernel contains $\mathbf{1}$.

Proof. To prove the existence and claimed properties of $\lim_{t \rightarrow \infty} (\Sigma^{[t]})^{-1}$, let $\{v_1, \dots, v_{d-1}\}$ be a basis of $\{\mathbf{1}\}^\perp$ in \mathbb{R}^d . Recall from the proof of Lemma S.5.3 that $\Sigma^{[t_0]}$ has rank $d - 1$, and let \tilde{v}_0 be its left null vector, i.e., $\tilde{v}_0^\top \Sigma^{[t_0]} = \mathbf{0}^\top$. Repeating the arguments in the proof of Lemma S.5.2 for $\check{S} = S^\top$ shows that $\tilde{v}_0^\top \mathbf{1} \neq 0$. Define the matrices

$W, \tilde{W} \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{d \times (d-1)}$ by

$$\begin{aligned} V &= (v_1, \dots, v_{d-1}) \\ W &= (v_0, v_1, \dots, v_{d-1}) \\ \tilde{W} &= (\tilde{v}_0, v_1, \dots, v_{d-1}) \end{aligned}$$

Note that $\Pi V = V$. Then

$$\begin{aligned} \tilde{W}^\top \mathbf{1} \mathbf{1}^\top W &= \begin{pmatrix} c & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{0} \mathbf{0}^\top \end{pmatrix}, \\ \tilde{W}^\top \Sigma^{[t_0]} W &= \begin{pmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & C \end{pmatrix}, \end{aligned}$$

for some constant $c \neq 0$ and $C \in \mathbb{R}^{(d-1) \times (d-1)}$ satisfying

$$\begin{aligned} C &= V^\top \Sigma^{[t_0]} V \\ &= V^\top \Pi \Sigma^{[t_0]} \Pi V \\ &= V^\top \Sigma V, \end{aligned}$$

which is symmetric positive definite, since Σ is symmetric positive semi-definite with kernel $\text{span}(\{\mathbf{1}\})$. Hence, using $\Sigma^{[t]} = (t - t_0) \mathbf{1} \mathbf{1}^\top + \Sigma^{[t_0]}$, we have

$$\begin{aligned} \Sigma^{[t]} &= (\tilde{W}^\top)^{-1} \begin{pmatrix} c(t - t_0) & \mathbf{0}^\top \\ \mathbf{0} & C \end{pmatrix} W^{-1}, \\ \implies (\Sigma^{[t]})^{-1} &= W \begin{pmatrix} c^{-1}(t - t_0)^{-1} & \mathbf{0}^\top \\ \mathbf{0} & C^{-1} \end{pmatrix} \tilde{W}^\top, \\ \implies \lim_{t \rightarrow \infty} (\Sigma^{[t]})^{-1} &= W \begin{pmatrix} 0 & \mathbf{0}^\top \\ \mathbf{0} & C^{-1} \end{pmatrix} \tilde{W}^\top. \end{aligned}$$

The latter matrix is symmetric since W and \tilde{W} differ only in the first column, and since $v_j \perp \mathbf{1}$ for all $j = 1, \dots, d-1$, its kernel contains $\mathbf{1}$. \square

Lemma S.5.5. *For $t > t_0$, let $\Theta^{[t]} := (\Sigma^{[t]})^{-1}$. Then*

$$\Sigma^+ = \lim_{t \rightarrow \infty} \Theta^{[t]}.$$

Proof. We will check the two criteria in Lemma S.1.4. Since $\Pi = I_d - d^{-1} \mathbf{1} \mathbf{1}^\top$ and since the vector $\mathbf{1}$ belongs to the kernel of the symmetric matrix $\lim_{t \rightarrow \infty} \Theta^{[t]}$, we have

$\Pi \lim_{t \rightarrow \infty} \Theta^{[t]} = \lim_{t \rightarrow \infty} \Theta^{[t]}$ and thus

$$\begin{aligned}
\Sigma \lim_{t \rightarrow \infty} \Theta^{[t]} &= \Pi S \Pi \lim_{t \rightarrow \infty} \Theta^{[t]} \\
&= \Pi \lim_{t \rightarrow \infty} S \Theta^{[t]} \\
&= \Pi \lim_{t \rightarrow \infty} (S + \mathbf{1}\mathbf{1}^\top t - \mathbf{1}\mathbf{1}^\top t) (S + \mathbf{1}\mathbf{1}^\top t)^{-1} \\
&= \Pi - \lim_{t \rightarrow \infty} \Pi \mathbf{1}\mathbf{1}^\top t (S + \mathbf{1}\mathbf{1}^\top t)^{-1} \\
&= \Pi.
\end{aligned}$$

Since $\Sigma = \Pi S \Pi$, it follows that $\text{Im } \Pi = \text{Im } \Sigma$ and thus that Π is the projection matrix onto the image of Σ . Further, the results from Lemma S.5.1 and Lemma S.5.4 yield

$$\ker \Sigma = \text{span}(\{\mathbf{1}\}) \subseteq \ker \lim_{t \rightarrow \infty} \Theta^{[t]}.$$

Hence, the claimed equality follows from Lemma S.1.4 with $A = \Sigma$ and $B = \lim_{t \rightarrow \infty} \Theta^{[t]}$. \square

Lemma S.5.6. *For Θ from Definition 3.2 it holds that*

$$\lim_{t \rightarrow \infty} (t\mathbf{1}\mathbf{1}^\top - \frac{1}{2}\Gamma)^{-1} = \Theta.$$

Proof. For $d = 2$, all valid variogram matrices are of the form $\Gamma_{11} = \Gamma_{22} = 0$ and $\Gamma_{12} = \Gamma_{21} =: \eta > 0$, and the expressions can directly be checked to be equal: we have $\Sigma^{(1)} = \Sigma^{(2)} = \gamma$ and

$$\begin{aligned}
(t\mathbf{1}\mathbf{1}^\top - \frac{1}{2}\Gamma)^{-1} &= \begin{bmatrix} t & t - \frac{1}{2}\eta \\ t - \frac{1}{2}\eta & t \end{bmatrix}^{-1} \\
&= \frac{1}{t^2 - (t - \frac{1}{2}\eta)^2} \begin{bmatrix} t & -t + \frac{1}{2}\eta \\ -t + \frac{1}{2}\eta & t \end{bmatrix} \\
&= \frac{1}{\frac{1}{2}\eta(2t - \frac{1}{2}\eta)} \begin{bmatrix} t & -t + \frac{1}{2}\eta \\ -t + \frac{1}{2}\eta & t \end{bmatrix} \\
&\rightarrow \frac{1}{\frac{1}{2}\eta} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1/\eta & -1/\eta \\ -1/\eta & 1/\eta \end{bmatrix} = \Theta, \quad t \rightarrow \infty.
\end{aligned}$$

For $d \geq 3$, write $V = \{1, \dots, d\}$, consider $\Sigma^{[t]} = t\mathbf{1}\mathbf{1}^\top - \frac{1}{2}\Gamma$ (which is positive definite for large enough t), let $Y^{[t]} \sim \mathcal{N}(0, \Sigma^{[t]})$, and let $Y^{[t,k]}$ be the random vector $Y^{[t]}$ conditioned on the event $\{Y_k^{[t]} = y_k\}$ for some $y_k \in \mathbb{R}$. Then the $(d-1)$ -dimensional

random vector $Y_{V \setminus \{k\}}^{[t,k]} = (Y_i^{[t,k]})_{i \in V \setminus \{k\}}$ is multivariate normal with covariance matrix $\Sigma^{[t,k]}$ given by

$$\begin{aligned}
\Sigma^{[t,k]} &= \Sigma_{\setminus k}^{[t]} - \Sigma_{\cdot, k}^{[t]} \left(\Sigma_{k, k}^{[t]} \right)^{-1} \Sigma_{k, \cdot}^{[t]} \\
&= \left(\mathbf{1}\mathbf{1}^\top t - \frac{1}{2} \Gamma_{\setminus k} \right) - \left(\mathbf{1}t - \frac{1}{2} \Gamma_{\cdot, k} \right) \frac{1}{t} \left(\mathbf{1}^\top t - \frac{1}{2} \Gamma_{k, \cdot} \right) \\
&= \mathbf{1}\mathbf{1}^\top t - \frac{1}{2} \Gamma_{\setminus k} - \mathbf{1}\mathbf{1}^\top t + \frac{1}{2} \mathbf{1} \Gamma_{k, \cdot} + \frac{1}{2} \Gamma_{\cdot, k} \mathbf{1}^\top + \frac{1}{4t} \Gamma_{\cdot, k} \Gamma_{k, \cdot} \\
&= \frac{1}{2} \left(\mathbf{1} \Gamma_{k, \cdot} + \Gamma_{\cdot, k} \mathbf{1}^\top - \Gamma_{\setminus k} \right) + o(1) \\
&= \Sigma^{(k)} + o(1), \quad t \rightarrow \infty,
\end{aligned}$$

with $\Sigma^{(k)}$ as in Definition 3.1. Since $\Sigma^{(k)}$ is positive definite, so is $\Sigma^{[t,k]}$ for sufficiently large t , and the respective precision matrices satisfy

$$\Theta^{[t,k]} := (\Sigma^{[t,k]})^{-1} \longrightarrow (\Sigma^{(k)})^{-1} =: \Theta^{(k)}, \quad t \rightarrow \infty.$$

Recall $\Theta^{[t]} = (\Sigma^{[t]})^{-1}$. Classical properties of the Schur complement of a block matrix (see, e.g. Rue and Held, 2005, Theorem 2.5) imply

$$\Theta^{[t,k]} = (\Theta^{[t]})_{(V \setminus \{k\}) \times (V \setminus \{k\})}$$

and hence

$$\Theta_{ij}^{[t]} \rightarrow \Theta_{ij}^{(k)}, \quad i, j \neq k, \quad t \rightarrow \infty.$$

Since Θ is defined by $\Theta_{ij} = \Theta_{ij}^{(k)}$, this implies

$$\lim_{t \rightarrow \infty} \left(t \mathbf{1}\mathbf{1}^\top - \frac{1}{2} \Gamma \right)^{-1} = \Theta. \quad \square$$

Proof of Proposition 3.3. Combining the results above yields

$$\begin{aligned}
\Theta &= \lim_{t \rightarrow \infty} \left(t \mathbf{1}\mathbf{1}^\top - \frac{1}{2} \Gamma \right)^{-1} && \text{(Lemma S.5.6)} \\
&= \left(\Pi \left(-\frac{1}{2} \Gamma \right) \Pi \right)^+ && \text{(Lemma S.5.5)} \\
&= \left(\Pi S \Pi \right)^+ && \text{(Lemma S.5.8)} \\
&= \lim_{t \rightarrow \infty} \left(t \mathbf{1}\mathbf{1}^\top + S \right)^{-1}. && \text{(Lemma S.5.5)}
\end{aligned}$$

Using the invertibility of the map that sends a matrix to its Moore–Penrose inverse, the equality $\Theta = \left(\Pi \left(-\frac{1}{2} \Gamma \right) \Pi \right)^+$ immediately implies that $\Pi S \Pi = \Pi \left(-\frac{1}{2} \Gamma \right) \Pi$ is a necessary condition for $\left(\Pi S \Pi \right)^+ = \Theta$. \square

Lemma S.5.7. Let $\Gamma \in \mathcal{D}_d$, $\Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi$, and $\Theta = \Sigma^+$. For $S \in \mathbb{R}^{d \times d}$, we have $\Pi S \Pi = \Sigma$ if and only if

$$\Theta S \Theta = \Theta. \quad (\text{S.5.1})$$

Proof. First, suppose $\Pi S \Pi = \Sigma = \Theta^+$. Since Π is the projection matrix onto the image of Σ and thus also onto the image of $\Theta = \Sigma^+$ (see proof of Lemma S.5.5),

$$\Theta S \Theta = \Theta \Pi S \Pi \Theta = \Theta \Theta^+ \Theta = \Theta.$$

Second, let (S.5.1) be satisfied. Since $\Theta^+ \Theta = \Theta \Theta^+$ is the projection matrix onto the image of Θ and thus equal to Π , we find

$$\Pi S \Pi = \Theta^+ \Theta S \Theta \Theta^+ = \Theta^+ \Theta \Theta^+ = \Theta^+ = \Sigma. \quad \square$$

Lemma S.5.8. Let $\Gamma \in \mathcal{D}_d$ and $\Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi$ and recall the map γ in (3.2). For symmetric $S \in \mathbb{R}^{d \times d}$, the identity $\Pi S \Pi = \Sigma$ is equivalent to

$$\gamma(S) = \Gamma.$$

Since all $\Gamma \in \mathcal{D}_d$ are symmetric and $\gamma(\cdot)$ preserves (a)symmetry, this Lemma also shows the following implication:

$$\gamma(S) = \Gamma \implies S = S^\top \wedge \Pi S \Pi = \Pi(-\frac{1}{2}\Gamma)\Pi.$$

Proof. Recall that

$$\gamma(S) = \mathbf{1} \operatorname{diag}(S)^\top + \operatorname{diag}(S) \mathbf{1}^\top - 2S.$$

On the one hand, if $\gamma(S) = \Gamma$, then, since $\Pi \mathbf{1} = \mathbf{0}$, we get

$$\Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi = \Pi(-\frac{1}{2}\gamma(S))\Pi = \Pi S \Pi.$$

On the other hand, for symmetric S and writing $v := S \mathbf{e}_d$ with $\mathbf{e}_d = d^{-1} \mathbf{1}$, we have, by $\Pi = I_d - \mathbf{e}_d \mathbf{1}^\top = I_d - \mathbf{1} \mathbf{e}_d^\top$ and the linearity of γ ,

$$\begin{aligned} \gamma(\Pi S \Pi) &= \gamma(\Pi(S - v \mathbf{1}^\top)) \\ &= \gamma(S - \mathbf{1} v^\top - v \mathbf{1}^\top + \mathbf{1} \mathbf{e}_d^\top v \mathbf{1}^\top) \\ &= \gamma(S) - \gamma(\mathbf{1} v^\top + v \mathbf{1}^\top) + \gamma(\mathbf{1} \mathbf{e}_d^\top v \mathbf{1}^\top) \\ &= \gamma(S), \end{aligned}$$

as $\operatorname{diag}(\mathbf{1} v^\top) = a = \operatorname{diag}(v \mathbf{1}^\top)$ and thus $\gamma(\mathbf{1} v^\top + v \mathbf{1}^\top) = \mathbf{0} \mathbf{0}^\top \in \mathbb{R}^{d \times d}$. It follows that $\Pi S \Pi = \Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi$ implies

$$\gamma(S) = \gamma(\Pi S \Pi) = \gamma(\Pi(-\frac{1}{2}\Gamma)\Pi) = \gamma(-\frac{1}{2}\Gamma) = \Gamma,$$

since $\operatorname{diag}(\Gamma) = \mathbf{0}$. □

S.5.2 Proof of Proposition 3.4

Definition S.5.9. *The covariance transform γ is defined for arbitrary square matrices as*

$$\begin{aligned}\gamma : \mathbb{R}^{d \times d} &\longrightarrow \mathbb{R}^{d \times d} \\ S &\longmapsto \mathbf{1} \operatorname{diag}(S)^\top + \operatorname{diag}(S) \mathbf{1}^\top - 2S.\end{aligned}$$

This definition does not require S to have any special properties but also does not guarantee many useful properties of $\gamma(S)$. A more useful mapping can be obtained by restricting the domain as follows.

Lemma S.5.10. *Recall \mathcal{D}_d from (3.1). Then $\gamma(\mathcal{R}) \subseteq \mathcal{D}_d$ for any \mathcal{R} satisfying*

$$\mathcal{R} \subseteq \left\{ \Sigma \in \mathbb{R}^{d \times d} : \Sigma = \Sigma^\top, \Sigma \text{ p.s.d.}, \ker \Sigma \cap \{\mathbf{1}\}^\perp = \{\mathbf{0}\} \right\}.$$

Proof. Let $\Gamma = \gamma(\Sigma)$ for some $\Sigma \in \mathcal{R}$. Then clearly $\Gamma = \Gamma^\top$ and $\operatorname{diag}(\Gamma) = \mathbf{0}$. Moreover, for $\mathbf{0} \neq v \in \{\mathbf{1}\}^\perp$, we have

$$v^\top \Gamma v = v^\top \gamma(\Sigma) v = v^\top (\mathbf{1} \operatorname{diag}(\Sigma)^\top + \operatorname{diag}(\Sigma) \mathbf{1}^\top - 2\Sigma) v = -2v^\top \Sigma v < 0,$$

implying that $\Gamma \in \mathcal{D}_d$. □

Proof of Proposition 3.4. Lemmas S.5.1 and S.5.10 show that σ and γ do indeed map between \mathcal{D}_d and \mathcal{P}_d^1 . Since both maps consist only of elementary matrix operations and the continuous map $\operatorname{diag}(\cdot)$, they are also continuous. Using the intermediate results from the proof of Lemma S.5.8, it can be seen that

$$\begin{aligned}\gamma(\sigma(\Gamma)) &= \gamma(\Pi(-\tfrac{1}{2}\Gamma)\Pi) = \Gamma, & \forall \Gamma \in \mathcal{D}_d, \\ \sigma(\gamma(\Sigma)) &= \Pi(-\tfrac{1}{2}\gamma(\Sigma))\Pi = \Sigma, & \forall \Sigma \in \mathcal{P}_d^1,\end{aligned}$$

implying that σ is bijective between \mathcal{D}_d and \mathcal{P}_d with continuous inverse $\sigma^{-1} = \gamma$.

By Lemma S.1.3, Item (4), the Moore–Penrose inverse of a symmetric positive semi-definite matrix is again symmetric positive semi-definite. Hence, θ also maps \mathcal{D}_d into \mathcal{P}_d^1 and can be written as the composition $\varphi \circ \sigma$, with

$$\begin{aligned}\varphi : \mathcal{P}_d^1 &\longrightarrow \mathcal{P}_d^1 \\ \Sigma &\longmapsto \Sigma^+.\end{aligned}$$

As shown in Rakočević (1997), the map φ is continuous, and hence is a continuous, idempotent bijection; note that it is surjective since every $\Sigma \in \mathcal{P}_d^1$ can be written as $\Sigma = \varphi(\varphi(\Sigma))$. Therefore, θ is also continuous with continuous inverse $\theta^{-1} = \gamma \circ \varphi$. □

S.5.3 Proofs of density expressions

Proof of Proposition 3.6. For $k \in \{1, \dots, d\}$, let $\lambda^{(k)}$ denote the density expression from Definition 3.1 with that specific choice of k . Extending the notation used there, let

$$\tilde{\mu}^{(k)} = \left(-\frac{1}{2}\Gamma_{ik}\right)_{i=1,\dots,d} \in \mathbb{R}^d.$$

Note that

$$(y - \mathbf{1}y_k - \tilde{\mu}^{(k)})_k = 0,$$

and hence $\lambda^{(k)}$ can be rewritten as

$$\lambda^{(k)}(y; \Gamma) = \frac{\exp(-y_k)}{\sqrt{(2\pi)^{d-1} |\Sigma^{(k)}|}} \exp\left(-\frac{1}{2}\|y - \mathbf{1}y_k - \tilde{\mu}^{(k)}\|_{\Theta}^2\right).$$

From Röttger et al. (2021, Lemma 4.4) it follows that the value of the normalizing constant is independent of the choice of k . Using $\Theta \mathbf{1} = \mathbf{0}$ and applying logarithms, the above equation becomes

$$\begin{aligned} \log \lambda^{(k)}(y; \Gamma) &= c_1 - y_k - \frac{1}{2}\|y - \tilde{\mu}^{(k)}\|_{\Theta}^2, \quad \text{where} \\ c_1 &= -\frac{1}{2} \log\left((2\pi)^{d-1} |\Sigma^{(1)}|\right). \end{aligned}$$

Next, let $v \in \mathbb{R}^d$ be such that $\mathbf{1}^\top v = 1$. Then, since the value of $\lambda^{(k)}$ is the same for all $k \in \{1, \dots, d\}$, we find

$$\begin{aligned} \log \lambda(y; \Gamma) &= \sum_{k=1}^d v_k \log \lambda^{(k)}(y; \Gamma) \\ &= \sum_{k=1}^d v_k \left(c_1 - y_k - \frac{1}{2}\|y - \tilde{\mu}^{(k)}\|_{\Theta}^2\right) \\ &= c_1 - v^\top y - \frac{1}{2} \sum_{k=1}^d v_k \|y - \tilde{\mu}^{(k)}\|_{\Theta}^2. \end{aligned} \tag{S.5.2}$$

The last sum can be rewritten as

$$\begin{aligned} \sum_{k=1}^d v_k \|y - \tilde{\mu}^{(k)}\|_{\Theta}^2 &= \|y\|_{\Theta}^2 + \sum_{k=1}^d v_k \langle y, \Gamma_{\cdot k} \rangle_{\Theta} + \sum_{k=1}^d v_k \left\| \frac{1}{2} \Gamma_{\cdot k} \right\|_{\Theta}^2 \\ &= \|y\|_{\Theta}^2 + \langle y, \Gamma v \rangle_{\Theta} + \left\| \frac{1}{2} \Gamma v \right\|_{\Theta}^2 + c_v \\ &= \left\| y + \frac{1}{2} \Gamma v \right\|_{\Theta}^2 + c_v, \end{aligned} \tag{S.5.3}$$

with

$$c_v = \left(\sum_{k=1}^d v_k \left\| \frac{1}{2} \Gamma \cdot k \right\|_{\Theta}^2 \right) - \left\| \frac{1}{2} \Gamma v \right\|_{\Theta}^2.$$

Moreover, $\Pi(-\frac{1}{2}\Gamma)\Pi = \Sigma = \Theta^+$. As $I_d = \Pi + \mathbf{e}_d \mathbf{1}^\top$ and $\mathbf{1}^\top v = 1$, we have $v = \Pi v + \mathbf{e}_d$ and thus

$$\begin{aligned} \Pi(-\tfrac{1}{2}\Gamma)v &= \Pi(-\tfrac{1}{2}\Gamma)(\Pi v + \mathbf{e}_d) \\ &= \Pi(-\tfrac{1}{2}\Gamma)\Pi v + \Pi(-\tfrac{1}{2}\Gamma)\mathbf{e}_d \\ &= \Theta^+ v + \Theta^+ \Theta(-\tfrac{1}{2}\Gamma)\mathbf{e}_d \\ &= \Theta^+(v + \Theta(-\tfrac{1}{2}\Gamma)\mathbf{e}_d) \\ &= \Theta^+(v + r_{\Theta}) \end{aligned}$$

yielding

$$\mu_v = \Pi(-\tfrac{1}{2}\Gamma)v. \quad (\text{S.5.4})$$

Since $\Pi\Theta = \Theta\Pi = \Theta\Theta^+\Theta = \Theta$, it follows from (S.5.4) that

$$\begin{aligned} \|y - \mu_v\|_{\Theta}^2 &= \|y - \Pi(-\tfrac{1}{2}\Gamma)v\|_{\Theta}^2 \\ &= \|y + \tfrac{1}{2}\Pi\Gamma v\|_{\Theta}^2 \\ &= \|y + \tfrac{1}{2}\Gamma v\|_{\Theta}^2. \end{aligned} \quad (\text{S.5.5})$$

Plugging the identities (S.5.3) and (S.5.5) back into the expression for $\log \lambda(y; \Gamma)$ in (S.5.2) yields (3.10):

$$\lambda(y; \Theta) = c_{\Theta, v} \cdot \exp(-y^\top v) \cdot \exp(-\tfrac{1}{2}\|y - \mu_v\|_{\Theta}^2),$$

with

$$c_{\Theta, v} = \exp(c_1 - \tfrac{1}{2}c_v).$$

To prove (3.9) let $v = \mathbf{e}_d$ and rearrange (3.10) as follows:

$$\begin{aligned} \lambda(y; \Theta) &= c_{\Theta, \mathbf{e}_d} \cdot \exp(-y^\top \mathbf{e}_d) \cdot \exp(-\tfrac{1}{2}\|y - \mu_{\mathbf{e}_d}\|_{\Theta}^2) \\ &= c_{\Theta, \mathbf{e}_d} \cdot \exp\left(-y^\top \mathbf{e}_d - \tfrac{1}{2}\|y + \tfrac{1}{2}\Gamma \mathbf{e}_d\|_{\Theta}^2\right) \\ &= c_{\Theta, \mathbf{e}_d} \cdot \exp\left(-\tfrac{1}{2}y^\top \Theta y - y^\top (\mathbf{e}_d + \tfrac{1}{2}\Theta \Gamma \mathbf{e}_d) - \tfrac{1}{2}\|\tfrac{1}{2}\Gamma \mathbf{e}_d\|_{\Theta}^2\right) \\ &= c_{\Theta} \cdot \exp\left(-\tfrac{1}{2}y^\top \Theta y - y^\top (\mathbf{e}_d - r_{\Theta})\right), \end{aligned}$$

with

$$c_{\Theta} = \exp\left(c_1 - \frac{1}{2}c_{\mathbf{e}_d} - \frac{1}{8}\|\Gamma\mathbf{e}_d\|_{\Theta}^2\right).$$

The expressions for c_{Θ} and $c_{\Theta,v}$ can be simplified further. First, since $\Theta^{(1)} = (\Sigma^{(1)})^{-1}$,

$$\exp(c_1) = (2\pi)^{-(d-1)/2}|\Sigma^{(1)}|^{-1/2} = (2\pi)^{-(d-1)/2}|\Theta^{(1)}|^{1/2}.$$

By Lemma S.5.11, the diagonal element $\|\frac{1}{2}\Gamma_{\cdot k}\|_{\Theta}^2 = ((-\frac{1}{2}\Gamma)\Theta(-\frac{1}{2}\Gamma))_{kk}$ does not depend on $k \in V$ and is equal to $c(\Gamma) = \|\frac{1}{2}\Gamma\mathbf{e}_d\|_{\Theta}^2 - \mathbf{e}_d^{\top}(-\frac{1}{2}\Gamma)\mathbf{e}_d$. It follows that

$$c_{\mathbf{e}_d} = -\mathbf{e}_d^{\top}(-\frac{1}{2}\Gamma)\mathbf{e}_d.$$

But then

$$\begin{aligned} c_{\Theta} &= (2\pi)^{-(d-1)/2}|\Theta^{(1)}|^{1/2} \exp\left(\frac{1}{2}\mathbf{e}_d^{\top}(-\frac{1}{2}\Gamma)\mathbf{e}_d - \frac{1}{2}\|\frac{1}{2}\Gamma\mathbf{e}_d\|_{\Theta}^2\right) \\ &= (2\pi)^{-(d-1)/2}|\Theta^{(1)}|^{1/2} \exp\left(-\frac{1}{2}c(\Gamma)\right). \end{aligned}$$

Second, for $c_{\Theta,v}$, note that, again by Lemma S.5.11,

$$c_v = c(\Gamma) - \|\frac{1}{2}\Gamma v\|_{\Theta}^2$$

and thus, since $\Theta\Pi = \Pi\Theta = \Theta$,

$$\begin{aligned} c_{\Theta,v} &= (2\pi)^{-(d-1)/2}|\Theta^{(1)}|^{1/2} \exp\left(-\frac{1}{2}c(\Gamma) + \frac{1}{2}\|\frac{1}{2}\Gamma v\|_{\Theta}^2\right) \\ &= c_{\Theta} \cdot \exp\left(\frac{1}{2}\|\frac{1}{2}\Gamma v\|_{\Theta}^2\right) \\ &= c_{\Theta} \cdot \exp\left(\frac{1}{2}\|\mu_v\|_{\Theta}^2\right). \end{aligned}$$

The latter identity can also be found by comparing (3.9) with (3.10): the terms not depending on y must be equal. \square

Lemma S.5.11. *Let $\Gamma \in \mathcal{D}_d$ and let $\Theta = (\Pi(-\frac{1}{2}\Gamma)\Pi)^+$. Then*

$$\begin{aligned} (-\frac{1}{2}\Gamma)\Theta(-\frac{1}{2}\Gamma) &= -\frac{1}{2}\Gamma + c(\Gamma)\mathbf{1}\mathbf{1}^{\top} \quad \text{where} \\ c(\Gamma) &= \mathbf{e}_d^{\top}(-\frac{1}{2}\Gamma)\Theta(-\frac{1}{2}\Gamma)\mathbf{e}_d - \mathbf{e}_d^{\top}(-\frac{1}{2}\Gamma)\mathbf{e}_d. \end{aligned}$$

Since Γ has a zero diagonal, this implies that $\Gamma\Theta\Gamma$ has a constant diagonal.

Proof. Obviously, it is sufficient to show that the matrix $(-\frac{1}{2}\Gamma)\Theta(-\frac{1}{2}\Gamma)$ has constant diagonal. Recall the definitions $\Sigma = \Pi(-\frac{1}{2}\Gamma)\Pi$, $\mathbf{e}_d = d^{-1}\mathbf{1}$ and $\Pi = I_d - \mathbf{1}\mathbf{e}_d^\top = I_d - \mathbf{e}_d\mathbf{1}^\top$. Since $\Pi\Theta = \Theta\Pi = \Theta$, we have

$$\begin{aligned} (-\tfrac{1}{2}\Gamma)\Theta(-\tfrac{1}{2}\Gamma) &= (\Pi + \mathbf{1}\mathbf{e}_d^\top)(-\tfrac{1}{2}\Gamma)\Pi\Theta(-\tfrac{1}{2}\Gamma) \\ &= \Pi(-\tfrac{1}{2}\Gamma)\Pi\Theta(-\tfrac{1}{2}\Gamma) + \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Pi\Theta(-\tfrac{1}{2}\Gamma) \\ &= \Pi(-\tfrac{1}{2}\Gamma) + \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta(-\tfrac{1}{2}\Gamma) \\ &= -\tfrac{1}{2}\Gamma - \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma) + \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta(-\tfrac{1}{2}\Gamma). \end{aligned}$$

Since $\Theta\Pi(-\frac{1}{2}\Gamma)\Pi = \Theta\Sigma = \Pi$, we have

$$\begin{aligned} \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta(-\tfrac{1}{2}\Gamma) &= \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta\Pi(-\tfrac{1}{2}\Gamma)(\Pi + \mathbf{e}_d\mathbf{1}^\top) \\ &= \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta\Pi(-\tfrac{1}{2}\Gamma)\Pi + \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta(-\tfrac{1}{2}\Gamma)\mathbf{e}_d\mathbf{1}^\top \\ &= \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Pi + \underbrace{(\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Theta(-\tfrac{1}{2}\Gamma)\mathbf{e}_d)}_{\text{scalar}} \mathbf{1}\mathbf{1}^\top \end{aligned}$$

with

$$\begin{aligned} \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\Pi &= \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma) - \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\mathbf{e}_d\mathbf{1}^\top \\ &= \mathbf{1}\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma) - \underbrace{(\mathbf{e}_d^\top(-\tfrac{1}{2}\Gamma)\mathbf{e}_d)}_{\text{scalar}} \mathbf{1}\mathbf{1}^\top. \end{aligned}$$

Adding up, we get the stated identity. \square

Proof of Corollary 3.7. A direct computation shows that Π_v is the inverse of Π when the latter is seen as a linear map from $\{v\}^\perp$ to $\{\mathbf{1}\}^\perp$; specifically, $\Pi(\Pi_v x) = x$ for $x \in \{\mathbf{1}\}^\perp$ and $\Pi_v(\Pi y) = y$ for $y \in \{v\}^\perp$. We have $W_v = \Pi_v X_v$ where the random vector $X_v \sim \mathcal{N}(\mu_v, \Theta^+)$ is concentrated on $\{\mathbf{1}\}^\perp$. The distribution of W_v can thus be interpreted as the one resulting from the invertible linear transformation Π_v from $\{\mathbf{1}\}^\perp$ onto $\{v\}^\perp$ applied to a normal random vector X_v on $\{\mathbf{1}\}^\perp$ with mean μ_v and covariance matrix Θ^+ . The matrix Π_v is a kind of oblique projection onto $\{v\}^\perp$, since the image of Π_v is $\{v\}^\perp$ and since $\Pi_v y = y$ for $y \in \{v\}^\perp$. In the special case $v = \mathbf{e}_d = d^{-1}\mathbf{1}$, we have $\{v\}^\perp = \{\mathbf{1}\}^\perp$ and $\Pi_v = \Pi$, so that $W_v = X_v$.

The density of $X_v \sim \mathcal{N}(\mu_v, \Theta^+)$ with respect to the $(d-1)$ -dimensional Lebesgue measure on $\{\mathbf{1}\}^\perp$ can be expressed as (e.g., Rao, 2002)

$$f_{X_v}(w) \propto \exp(-\tfrac{1}{2}\|w - \mu_v\|_{\Theta}^2), \quad w \in \{\mathbf{1}\}^\perp.$$

We have $W_v = \Pi_v X_v$ where Π_v is a linear bijection from $\{\mathbf{1}\}^\perp$ onto $\{v\}^\perp$ with inverse Π . As a consequence, the density of W_v with respect to the $(d-1)$ -dimensional Lebesgue measure on $\{v\}^\perp$ satisfies, for $w \in \{v\}^\perp$,

$$\begin{aligned} f_{W_v}(w) &\propto f_{X_v}(\Pi w) \\ &\propto \exp\left(-\frac{1}{2}\|\Pi w - \mu_v\|_\Theta^2\right) \\ &= \exp\left(-\frac{1}{2}\|w - \mu_v\|_\Theta^2\right), \end{aligned}$$

as $\Pi\Theta = \Theta\Pi = \Theta$. Furthermore, the density of $R \sim \text{Exp}(1)$ is $f_R(r) = \exp(-r)$, $r > 0$. Observe that the mapping $(w, r) \mapsto y = w + r\mathbf{1}$ is a linear bijection from $\{v\}^\perp \times \mathbb{R}$ to \mathbb{R}^d with inverse $y \mapsto (\Pi_v y, v^\top y)$ from \mathbb{R}^d to $\{v\}^\perp \times \mathbb{R}$: indeed, we have $(w, r) = (\Pi_v(w + r\mathbf{1}), (w + r\mathbf{1})^\top v)$ for $(w, r) \in \{v\}^\perp \times \mathbb{R}$ and $y = \Pi_v y + \mathbf{1}v^\top y$ for $y \in \mathbb{R}^d$. Hence, the probability density function of $Y_v = W_v + R\mathbf{1}$ is proportional to the product of the densities of W_v and R : For $y \in \mathbb{R}^d$, we have

$$\begin{aligned} f_{Y_v}(y) &\propto f_R(v^\top y) \cdot f_{W_v}(\Pi_v y) \\ &\propto \exp(-v^\top y) \cdot \exp\left(-\frac{1}{2}\|\Pi_v y - \mu_v\|_\Theta^2\right) \\ &\propto \exp(-v^\top y) \cdot \exp\left(-\frac{1}{2}\|y - \mu_v\|_\Theta^2\right), \end{aligned}$$

the last step following from $\Theta\Pi_v = \Theta - \Theta\mathbf{1}v^\top = \Theta$ and similarly $\Pi_v^\top\Theta = \Theta$, since $\Theta\mathbf{1} = \mathbf{0}$ by assumption. The density of Y_v is proportional to the one in (3.10), and, since $v \geq \mathbf{0}$, the support of Y_v is the same as the one of Y from Corollary 3.7, conditioned on the event $\{v^\top Y > 0\}$. Hence, the two random vectors are equal in distribution. \square

S.5.4 Proofs of matrix completion results

S.5.4.1 Decomposable graphs

Lemma S.5.12 (Bakonyi and Woerdeman (2011), Theorem 2.2.3). *The positive definite block matrix completion problem*

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} & ?? \\ \Sigma_{BA} & \Sigma_{BB} & \Sigma_{BC} \\ ?? & \Sigma_{CB} & \Sigma_{CC} \end{pmatrix}, \quad (\Sigma^{-1})_{AC} = 0 \quad (\text{S.5.6})$$

has the unique solution $\Sigma_{AC} = \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BC}$, and $\Sigma_{AC} = 0$ in the case of an empty separator $B = \emptyset$. Here, “??” denotes a matrix of adequate size with all entries “?”.

Lemma S.5.13. *Let $G = (V, E)$, for $V = \{1, \dots, d\}$, be an undirected graph, with some node $k \in V$ being connected to all other nodes. Let $\mathring{\Gamma}$ be a matrix specified on G that is conditionally negative definite on all fully specified principal submatrices. Let $\mathring{\Sigma}^{(k)} = \varphi_k(\mathring{\Gamma})$ for φ_k in Definition 3.1, for ease of notation indexed by $V \setminus \{k\}$.*

Then $\mathring{\Sigma}^{(k)}$ is positive definite on all fully specified principal submatrices and preserves specified entries in the sense that

$$\mathring{\Gamma}_{ij} \neq \text{“?”} \iff \mathring{\Sigma}_{ij}^{(k)} \neq \text{“?”}, \quad \forall i, j \neq k.$$

Proof. To show that specified entries are preserved, recall that

$$\mathring{\Sigma}_{ij}^{(k)} = \frac{1}{2} \left(\mathring{\Gamma}_{ik} + \mathring{\Gamma}_{jk} - \mathring{\Gamma}_{ij} \right), \quad i, j \neq k.$$

Since k is connected to all other nodes in G , $\mathring{\Gamma}_{ik}$ is specified for all $i \in V \setminus \{k\}$, yielding the claimed equivalence.

To show positive definiteness of fully specified submatrices, let M be the index set of such a submatrix and observe

$$\mathring{\Sigma}_{M,M}^{(k)} = \left(\varphi_k(\mathring{\Gamma}) \right)_{M,M} = \varphi_k \left(\mathring{\Gamma}_{M \cup \{k\}, M \cup \{k\}} \right),$$

which is positive definite. □

Proof of Lemma 4.2. First, note that φ_k is in fact bijective with inverse

$$\varphi_k^{-1} : \Sigma^{(k)} \mapsto \left(\tilde{\Sigma}_{ii}^{(k)} + \tilde{\Sigma}_{jj}^{(k)} - 2\tilde{\Sigma}_{ij}^{(k)} \right)_{i,j=1,\dots,d},$$

where $\tilde{\Sigma} \in \mathbb{R}^{d \times d}$ is defined as $\Sigma^{(k)}$ with a zero-valued k th row and column added. Upon a permutation of the indices, the $(d-1) \times (d-1)$ matrix $\overset{\circ}{\Sigma}^{(k)}$ takes the form (S.5.6) with blocks of indices $A = C_1 \setminus \{k\}$, $B = D_2 \setminus \{k\}$ (possibly empty), and $C = C_2 \setminus \{k\}$. Lemma S.5.12 permits to find the unique positive semi-definite completion of $\overset{\circ}{\Sigma}^{(k)}$, say $\Sigma^{(k)}$. The matrix $\Gamma := \varphi_k^{-1}(\Sigma^{(k)})$ is thus well-defined.

Since φ_k and φ_k^{-1} preserve specified entries, the condition

$$\Gamma_{ij} = \overset{\circ}{\Gamma}_{ij}, \quad \forall (i, j) \in \overline{E},$$

is satisfied by construction. Furthermore, from Definition 3.2 and Lemma S.5.12 it follows that

$$\Theta_{ij} = \Theta_{ij}^{(k)} = 0, \quad \forall (i, j) \notin \overline{E}.$$

The uniqueness of this completion follows from Corollary S.5.20 below. \square

Lemma S.5.14. *Let Y be a random variable following a multivariate generalized Pareto distribution with positive, continuous exponent measure density λ . Let $G = (V, E)$ be a connected, decomposable graph $G = (V, E)$, consisting of two cliques C_1, C_2 separated by $D_2 = C_1 \cap C_2$. Let $G' = (V', E')$ be a connected, decomposable graph with $V' = C_1$ and $E' \supseteq E|_{D_2}$. Suppose Y satisfies the (extremal) pairwise Markov property relative to G , and the marginal Y_{C_1} conditionally on $\max_{i \in C_1} Y_i > 0$ satisfies the (extremal) pairwise Markov property relative to G' .*

Then Y satisfies the (extremal) pairwise Markov property relative to the graph $G'' = (V, E'')$ with $E'' = E|_{C_2} \cup E'$.

Remark S.5.15. Since the pairwise Markov property only requires

$$(i, j) \notin E \implies Y_i \perp_e Y_j \mid Y_{\setminus \{i, j\}},$$

but not vice versa, a distribution can satisfy this for a number of distinct graphs. In particular, adding edges to a graph strictly weakens the condition above.

Proof. Since all graphs involved in the lemma are connected and decomposable, Theorem 1 from Engelke and Hitz (2020) can be used to show

$$\begin{aligned} \lambda(y) &= \frac{\lambda_{C_1}(y_{C_1})\lambda_{C_2}(y_{C_2})}{\lambda_{D_2}(y_{D_2})} \\ &= \frac{\lambda_{C_2}(y_{C_2}) \prod_{C \in \mathcal{C}'} \lambda_C(y_C)}{\lambda_{D_2}(y_{D_2}) \prod_{D \in \mathcal{D}'} \lambda_D(y_D)} \\ &= \frac{\prod_{C \in \mathcal{C}''} \lambda_C(y_C)}{\prod_{D \in \mathcal{D}''} \lambda_D(y_D)}, \end{aligned}$$

with \mathcal{C}'' , \mathcal{D}'' being the cliques and separators of G'' , and \mathcal{C}' , \mathcal{D}' those of G' . Here, the function λ_I , for non-empty $I \subset V$, is the exponent measure density corresponding to the I -th marginal Y_I conditionally on $\max_{i \in I} Y_i > 0$. By the same theorem, the above decomposition of $\lambda(y)$ implies that Y satisfies the pairwise Markov property relative to G'' . \square

Proof of Proposition 4.3. In order to formalize the procedure described in the Proposition, let $G = (V, E)$ be a connected decomposable graph and $\mathring{\Gamma}$ a partially conditionally negative definite matrix specified on G . Let $\mathcal{C} = \{C_1, \dots, C_N\}$ be the cliques of G , ordered according to the running intersection property (Section S.2), and w.l.o.g. assume that the vertices in V are ordered accordingly, in the sense $i < j$ for all $i \in C_k$ and $j \in C_l$ with $k < l$.

Let $\Gamma_1 = \mathring{\Gamma}|_{C_1}$, $J_n = C_1 \cup \dots \cup C_n$, and $K_n = J_n \times J_n$. Further, for $n = 2, \dots, N$, iteratively define

$$\begin{aligned} \mathring{\Gamma}_n &\in \mathring{\mathbb{R}}^{|J_n| \times |J_n|}, && \text{with entries} \\ (\mathring{\Gamma}_n)_{ij} &= \begin{cases} (\Gamma_{n-1})_{ij} & (i, j) \in \bar{K}_{n-1}, \\ \mathring{\Gamma}_{ij} & (i, j) \in \bar{E}_n \setminus \bar{K}_{n-1}, \\ \text{"?"} & \text{otherwise,} \end{cases} \\ \Gamma_n &= \mathfrak{C}(\mathring{\Gamma}_n) \in \mathbb{R}^{|J_n| \times |J_n|}, \end{aligned}$$

where \mathfrak{C} denotes the completion from Lemma 4.2. Setting $\Gamma = \Gamma_N$, the condition $\Gamma_{ij} = \mathring{\Gamma}_{ij}$ for all $(i, j) \in \bar{E}$ is satisfied by construction. The condition $\Theta_{ij} = 0$ for all $(i, j) \notin \bar{E}$ is satisfied too, since Lemma S.5.14 can be applied in each step. The uniqueness of this completion follows from Corollary S.5.20. \square

S.5.4.2 General graphs

The proofs shown here closely follow the proofs in Speed and Kiiveri (1986) and are adjusted where necessary to hold for conditionally negative definite variogram matrices instead of positive definite covariance matrices. Recall the map $\sigma(\cdot)$ in (3.7).

Definition S.5.16. Let $\Gamma_1, \Gamma_2 \in \mathcal{D}_d$ and $\Sigma_1 = \sigma(\Gamma_1)$, $\Sigma_2 = \sigma(\Gamma_2)$. Let P_1, P_2 denote the probability measures of two degenerate normally distributed random vectors with mean $\mathbf{0}$ and covariance matrices $\Sigma_1, \Sigma_2 \in \mathcal{P}_d^{\mathbf{1}}$, and let p_1, p_2 be the corresponding densities on $\{\mathbf{1}\}^\perp$ with respect to the $(d-1)$ -dimensional Lebesgue measure. The Kullback–Leibler divergence \mathcal{I} of these matrices is defined as

$$\mathcal{I}(\Gamma_1|\Gamma_2) := \mathcal{I}(\Sigma_1|\Sigma_2) := \mathcal{I}(P_1|P_2) = \mathbb{E}_{P_1}(\log p_1(x) - \log p_2(x)).$$

Lemma S.5.17. *With $\Gamma_i, \Sigma_i, i = 1, 2$ as above, and $|\cdot|_+$ denoting the pseudo-determinant (see Section S.1.2), we have*

$$\mathcal{I}(\Gamma_1|\Gamma_2) = -\frac{1}{2}\left(\log|\Sigma_2^+\Sigma_1|_+ + d - 1 - \text{tr}(\Sigma_2^+\Sigma_1)\right).$$

Proof. Since $\mathbb{E}_{P_1}(xx^\top) = \Sigma_1$ and since $p_i(x) \propto |\Sigma_i|_+^{-1/2} \exp(-\frac{1}{2}x^\top\Sigma_i^+x)$ with proportionality constant not depending on $i \in \{1, 2\}$, we have

$$\begin{aligned} \mathcal{I}(P_1|P_2) &= \mathbb{E}_{P_1}(\log p_1(x) - \log p_2(x)) \\ &= \frac{1}{2}\mathbb{E}_{P_1}\left(-\log|\Sigma_1|_+ - x^\top\Sigma_1^+x + \log|\Sigma_2|_+ + x^\top\Sigma_2^+x\right) \\ &= \frac{1}{2}\log\frac{|\Sigma_2|_+}{|\Sigma_1|_+} + \frac{1}{2}\mathbb{E}_{P_1}\left(-x^\top\Sigma_1^+x + x^\top\Sigma_2^+x\right) \\ &= -\frac{1}{2}\log|\Sigma_2^+\Sigma_1|_+ + \frac{1}{2}\mathbb{E}_{P_1}\left(-\text{tr}(\Sigma_1^+xx^\top) + \text{tr}(\Sigma_2^+xx^\top)\right) \\ &= -\frac{1}{2}\log|\Sigma_2^+\Sigma_1|_+ - \frac{1}{2}\text{tr}(\Sigma_1^+\Sigma_1) + \frac{1}{2}\text{tr}(\Sigma_2^+\Sigma_1). \end{aligned}$$

The statement then follows since $\Sigma_1^+\Sigma_1$ is equal to the projection matrix Π onto $\{\mathbf{1}\}^\perp$ and the trace of a projection matrix is equal to its rank. \square

Lemma S.5.18. *Consider a fixed $\Gamma_2 \in \mathcal{D}_d$. Then the function $\mathcal{I}(\cdot|\Gamma_2)$ is convex on \mathcal{D}_d .*

Proof. Since $\text{tr}(\cdot)$ and $\sigma(\cdot)$ are linear maps, it remains to be shown that, for a fixed $\Sigma_2 \in \mathcal{P}_d^1$, the map $A \mapsto \log|\Sigma_2^+A|_+$ is concave on \mathcal{P}_d^1 . Since Σ_2^+ and A are symmetric with the same kernel, Lemma S.1.7 shows that

$$\log|\Sigma_2^+A|_+ = \log|\Sigma_2^+|_+ + \log|A|_+,$$

which is a concave function in A if and only if $A \mapsto \log|A|_+$ is concave.

To show this, recall the notation from Lemma S.1.7 with $V = \text{span}(\{\mathbf{1}\})$, and let $s \in [0, 1]$, and $A_1, A_2 \in \mathcal{Q}_V$ positive semi-definite. Then indeed

$$\begin{aligned} \log|sA_1 + (1-s)A_2|_+ &= \log|sB'_1 + (1-s)B'_2| \\ &\geq s\log|B'_1| + (1-s)\log|B'_2| \\ &= s\log|A_1|_+ + (1-s)\log|A_2|_+, \end{aligned}$$

where the inequality follows from the log-concavity of the determinant for positive definite matrices (see e.g. Boyd and Vandenberghe, 2004, p. 73f). \square

In the following results, let $S \subseteq V \times V$ be such that $(i, j) \in S$ implies $(j, i) \in S$, i.e., S corresponds to the edges of an undirected graph on V , optionally extended by some pairs (i, i) , for $i \in V$.

Lemma S.5.19. *For $\Gamma_i, \Gamma'_i \in \mathcal{D}_d$, $\Sigma_i = \sigma(\Gamma_i)$, $\Theta_i = \theta(\Gamma_i)$, with $i \in \mathbb{N}$, the divergence $\mathcal{I}(\cdot | \cdot)$ has the following properties.*

(1) $\mathcal{I}(\Gamma_1 | \Gamma_2) \geq 0$, with equality if and only if $\Gamma_1 = \Gamma_2$.

(2) Given $\Gamma_1, \Gamma_3 \in \mathcal{D}_d$, if there exists a $\Gamma_2 \in \mathcal{D}_d$ such that

(a) $(\Gamma_2)_{ij} = (\Gamma_1)_{ij}$, for $(i, j) \in S$, and

(b) $(\Theta_2)_{ij} = (\Theta_3)_{ij}$, for $(i, j) \notin S$,

then

$$\mathcal{I}(\Gamma_1 | \Gamma_3) = \mathcal{I}(\Gamma_1 | \Gamma_2) + \mathcal{I}(\Gamma_2 | \Gamma_3).$$

If such a Γ_2 exists, it is unique.

(3) If $\{\Gamma_n\}$ and $\{\Gamma'_n\}$ are sequences contained in compact subsets of \mathcal{D}_d , then $\mathcal{I}(\Gamma_n | \Gamma'_n) \rightarrow 0$ implies $\Gamma_n - \Gamma'_n \rightarrow 0$ as $n \rightarrow \infty$.

Proof. Statement (1) is a well-known property of the Kullback–Leibler divergence, which can be applied since σ is injective (see Proposition 3.4), and each $\Sigma \in \mathcal{P}_d^1$ characterizes a different distribution.

To show statement (2), use the multiplicative property of the pseudo-determinant on \mathcal{P}_d^1 to compute

$$\begin{aligned} & -2(\mathcal{I}(\Gamma_1 | \Gamma_2) + \mathcal{I}(\Gamma_2 | \Gamma_3)) \\ &= \log |\Sigma_2^+ \Sigma_1|_+ + \text{tr}(\Sigma_1^+ \Sigma_1) - \text{tr}(\Sigma_2^+ \Sigma_1) + \log |\Sigma_3^+ \Sigma_2|_+ + \text{tr}(\Sigma_2^+ \Sigma_2) - \text{tr}(\Sigma_3^+ \Sigma_2) \\ &= \log |\Sigma_3^+ \Sigma_2 \Sigma_2^+ \Sigma_1|_+ + \text{tr}(\Sigma_1^+ \Sigma_1 - \Sigma_2^+ \Sigma_1 + \Sigma_2^+ \Sigma_2 - \Sigma_3^+ \Sigma_2) \\ &= \log |\Sigma_3^+ \Sigma_1|_+ + \text{tr}((\Sigma_2^+ - \Sigma_3^+)(\Sigma_2 - \Sigma_1) - \Sigma_3^+ \Sigma_1 + \Sigma_1^+ \Sigma_1) \\ &= -2\mathcal{I}(\Gamma_1 | \Gamma_3) + \text{tr}((\Sigma_2^+ - \Sigma_3^+)(\Sigma_2 - \Sigma_1)). \end{aligned}$$

Since $\Sigma_i = \Pi(-\frac{1}{2}\Gamma_i)\Pi$ and since $\Pi\Theta_i = \Theta_i\Pi = \Theta_i$, the invariance of the trace operator under cyclic permutations permits writing

$$\begin{aligned} \text{tr}((\Sigma_2^+ - \Sigma_3^+)(\Sigma_2 - \Sigma_1)) &= -\frac{1}{2}\text{tr}((\Theta_2 - \Theta_3)\Pi(\Gamma_2 - \Gamma_1)\Pi) \\ &= -\frac{1}{2}\text{tr}((\Theta_2 - \Theta_3)(\Gamma_2 - \Gamma_1)) \\ &= -\frac{1}{2} \sum_{(i,j) \in V \times V} (\Theta_2 - \Theta_3)_{ij} (\Gamma_2 - \Gamma_1)_{ij}. \end{aligned}$$

Each term in the sum is zero by (2)a and (2)b.

Uniqueness can be shown as in the positive definite case by considering Γ_2, Γ'_2 that satisfy (2)a and (2)b. If Γ_2, Γ'_2 both satisfy the said properties with respect to Γ_1 and Γ_3 , then they obviously also satisfy those properties with respect to $\Gamma'_1 = \Gamma'_3 = \Gamma_2$. But then, property (1) and the first part of property (2) yield

$$\begin{aligned} 0 &= \mathcal{I}(\Gamma_2|\Gamma_2) = \mathcal{I}(\Gamma_2|\Gamma'_2) + \mathcal{I}(\Gamma'_2|\Gamma_2) \\ \implies \mathcal{I}(\Gamma_2|\Gamma'_2) &= \mathcal{I}(\Gamma'_2|\Gamma_2) = 0 \\ \implies \Gamma_2 &= \Gamma'_2. \end{aligned}$$

Lastly, suppose $\{\Gamma_n\}$ and $\{\Gamma'_n\}$ are as in statement (3), but $\Gamma_n - \Gamma'_n \not\rightarrow 0$. Then there are convergent subsequences $\Gamma_{n_i} \rightarrow \Gamma^*$ and $\Gamma'_{n_i} \rightarrow \Gamma^{*'}$ with $\Gamma^* \neq \Gamma^{*'}$. By continuity of \mathcal{I} , $\mathcal{I}(\Gamma_{n_i} | \Gamma'_{n_i}) \rightarrow \mathcal{I}(\Gamma^* | \Gamma^{*'}) \neq 0$, which is a contradiction. \square

Corollary S.5.20. *Let $G = (V, E)$ be a connected graph, $\mathring{\Gamma}$ a partially conditionally negative definite matrix on G , and Γ a graphical completion of $\mathring{\Gamma}$, in the sense of Definition 4.1. Then Lemma S.5.19, Item (2), with $\Gamma_1 = \Gamma_3 = \Gamma$ and $S = \overline{E}$ shows the uniqueness of the completion from Lemma 4.2 and Propositions 4.3 and 4.5.*

Recall $S \subseteq V \times V$ as introduced prior to Lemma S.5.19.

Lemma S.5.21. *Let $S_1, \dots, S_m \subseteq S$ be such that their union is S . Let $\Gamma \in \mathcal{D}_d$, $\Theta \in \mathcal{P}_d^1$ (not necessarily $\Theta = \theta(\Gamma)$), and $\Gamma_n \in \mathcal{D}_d$, $\Theta_n = \theta(\Gamma_n)$, for $n \in \mathbb{N}_0$, with $\Gamma_0 = \Gamma$.*

If each Γ_n , $n \geq 1$, satisfies

$$\begin{aligned} (\Theta_n)_{ij} &= \Theta_{ij} && \text{if } (i, j) \in S_t, \\ (\Gamma_{n-1})_{ij} &= (\Gamma_n)_{ij} && \text{if } (i, j) \notin S_t, \end{aligned}$$

with $t \equiv n \pmod{m}$, then the sequence $(\Gamma_n)_n$ converges to the unique $\Gamma^S \in \mathcal{D}_d$ that, writing $\Theta^S = \theta(\Gamma^S)$, satisfies

$$\Theta_{ij}^S = \Theta_{ij} \quad \text{if } (i, j) \in S, \tag{S.5.7}$$

$$\Gamma_{ij}^S = \Gamma_{ij} \quad \text{if } (i, j) \notin S. \tag{S.5.8}$$

Proof. Using Lemma S.5.19, Item (2), decompose the Kullback–Leibler divergence $\mathcal{I}(\Gamma_0|\gamma(\Theta^+))$ as follows:

$$\begin{aligned} \mathcal{I}(\Gamma_{n-1}|\gamma(\Theta^+)) &= \mathcal{I}(\Gamma_{n-1}|\Gamma_n) + \mathcal{I}(\Gamma_n|\gamma(\Theta^+)) \\ \implies \mathcal{I}(\Gamma_0|\gamma(\Theta^+)) &= \mathcal{I}(\Gamma_n|\gamma(\Theta^+)) + \sum_{k=1}^n \mathcal{I}(\Gamma_{k-1}|\Gamma_k). \end{aligned}$$

From here it follows that $\sum_{k=1}^{\infty} \mathcal{I}(\Gamma_{k-1}|\Gamma_k)$ is a convergent series so that $\mathcal{I}(\Gamma_{k-1}|\Gamma_k) \rightarrow 0$ for $k \rightarrow \infty$, and also that

$$\{\Gamma_k\}_k \subseteq \{\Gamma' \in \mathcal{D}_d : \mathcal{I}(\Gamma'|\gamma(\Theta^+)) \leq \mathcal{I}(\Gamma_0|\gamma(\Theta^+))\} =: A,$$

which is a compact set due to the convexity of $\mathcal{I}(\Gamma|\gamma(\Theta^+))$ as a function of Γ (see Lemma S.5.18) and the facts that \mathcal{D}_d is a convex cone and $\Gamma \mapsto \mathcal{I}(\Gamma|\gamma(\Theta^+))$ attains its minimum uniquely at $\Gamma = \gamma(\Theta^+)$.

Hence, the sequence $v_s := (\Gamma_{sm+1}, \dots, \Gamma_{sm+m})$, $s \in \mathbb{N}$, has a convergent subsequence indexed by $s \in \mathbb{N}^* \subseteq \mathbb{N}$ with limit $(\Gamma_1^*, \dots, \Gamma_m^*)$. For any $2 \leq t \leq m$, the entries of this limit satisfy

$$(\Gamma_t^* - \Gamma_{t-1}^*) = (\Gamma_t^* - \Gamma_{sm+t}) + (\Gamma_{sm+t} - \Gamma_{sm+t-1}) + (\Gamma_{sm+t-1} - \Gamma_{t-1}^*).$$

For $s \in \mathbb{N}^* \rightarrow \infty$, the first and last term in this sum converge to zero by the definition of Γ_t^* , and the second term converges to zero by Lemma S.5.19, Item (3). Hence, $\Gamma_t^* = \Gamma_{t-1}^* =: \Gamma^*$ for all $2 \leq t \leq m$. Since the elements of the sequence Γ_{sm+t} , $s \in \mathbb{N}^*$, $t = 1, \dots, m$, satisfy (S.5.8) (always), and (S.5.7) infinitely often, the limit Γ^* satisfies both conditions, as well.

The above argument can be repeated for all other convergent subsequences of $(v_s)_{s \in \mathbb{N} \setminus \mathbb{N}^*}$ and Lemma S.5.19, Item (2), then shows that the limits of all these subsequences must be identical, hence $(\Gamma_k)_{k \in \mathbb{N}}$ converges to the unique Γ^S satisfying Eqs. (S.5.7) and (S.5.8). \square

Proof of Proposition 4.5. The result follows from Lemma S.5.21 with $S_i = E_i^c$ and Θ being the Laplacian matrix of the graph G (see Definition S.2.6). Uniqueness follows from Corollary S.5.20. \square

S.5.4.3 Example of non-completable graph

Proof of Example 4.7. A useful property of variogram matrices is the fact that they can equivalently be interpreted as Euclidean distance matrices. For instance, Gower (1982) shows that a matrix is (strictly) conditionally negative definite if and only if it is the $d \times d$ matrix consisting of the squared distances $\|p_i - p_j\|^2$ of a set of points $p_1, \dots, p_d \in \mathbb{R}^{d-1}$ that do not lie in a lower dimensional affine hyper-plane.

If there was a completion of the matrix $\mathring{\Gamma}$ from the Example, then the above interpretation as a (partial) Euclidean distance matrix would imply the existence of a set of points p_1, \dots, p_d , with distances $\|p_i - p_{i+1}\| = 1$ for $i = 1, \dots, d-1$, and

$\|p_1 - p_d\| = 2d$. However, by the triangle inequality this leads to the contradiction

$$2d = \|p_1 - p_d\| \leq \sum_{i=1}^{d-1} \|p_i - p_{i+1}\| = d.$$

Bakonyi and Johnson (1995) show that such a counter-example can be constructed for any non-decomposable graph. \square

S.5.4.4 Matrix completion as likelihood optimization

Proof of Proposition 5.1. Let $U : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d(d-1)/2}$ denote the mapping that maps a matrix to the vector containing the entries in its upper triangular part (excluding the diagonal). Note that the restriction $U|_{\mathcal{P}_d}$ is invertible, since the lower triangular part of $\Theta \in \mathcal{P}_d$ is defined by symmetry and the diagonal is such that the row sums are zero. In the computations below, we consider $P \in U(\mathcal{P}_d)$ and write $\Theta := \Theta(P)$ for notational convenience.

Consider the function $f(P) = \log |\Theta|_+ + \frac{1}{2} \text{tr}(\bar{\Gamma}\Theta)$. Röttger et al. (2021, Proposition A.5) show that

$$\nabla_P \log |\Theta|_+ = U(-\gamma(\Theta^+)).$$

Furthermore, by symmetry of the matrices Θ and $\bar{\Gamma}$,

$$\nabla_P \text{tr}(\Theta\bar{\Gamma}) = \nabla_P \sum_{i,j \in 1, \dots, d} \Theta_{ij} \bar{\Gamma}_{ij} = \nabla_P \sum_{\substack{i,j \in 1, \dots, d \\ i < j}} 2\Theta_{ij} \bar{\Gamma}_{ij} = U(2\bar{\Gamma}),$$

and hence

$$\nabla_P f(\Theta) = U(\bar{\Gamma} - \gamma(\Theta)).$$

The map $\Theta \mapsto \text{tr}(\bar{\Gamma}\Theta)$ is linear and the proof of Lemma S.5.18 shows that the map $\Theta \mapsto \log |\Theta|_+$ is concave. Hence, the maximizer of f under the constraint $\Theta_{ij} = 0$ for $(i, j) \notin \bar{E}$ satisfies

$$\gamma(\Theta)_{ij} = \bar{\Gamma}_{ij}, \quad \forall (i, j) \in \bar{E}.$$

Proposition 4.5 shows that the unique solution satisfying this condition and the constraint is $\theta(\mathfrak{C}_G(\bar{\Gamma}))$. \square

S.5.4.5 Matrix completion is continuous

Proof of Lemma 5.2. Enumerate the edges as $E = \{e_1, \dots, e_m\}$ with $m = |E|$ the cardinality of E . Let θ denote the mapping sending a variogram matrix Γ to its precision matrix $\Theta = \theta(\Gamma)$ as constructed in Proposition 3.3. Let π_G be the restriction map

$$\begin{aligned} \pi_G : \mathbb{R}^{d \times d} &\longrightarrow \mathbb{R}^m \\ M &\longmapsto v \end{aligned}$$

where v has entries v_k for $k = 1, \dots, m$, with

$$v_k = M_{ij} \text{ for } e_k = (i, j), i < j.$$

In words, π_G extracts from a matrix M the m elements M_{ij} corresponding to the edges $(i, j) \in E$ and stacks them in a vector.

Let \mathcal{Q}_G be the set of symmetric $d \times d$ matrices with zero row sums and with zeros in off-diagonal positions corresponding to non-edges of G :

$$\mathcal{Q}_G := \{M \in \mathbb{R}^{d \times d} : M = M^\top, M\mathbf{1} = \mathbf{0}, M_{ij} = 0 \forall (i, j) \notin \overline{E}\}.$$

Clearly, $M \in \mathcal{Q}_G$ is determined by its elements M_{ij} for $(i, j) \in E$ such that $i < j$, i.e., by $\pi_G(M)$. Indeed, the elements of M below the diagonal are determined by the symmetry constraint $M = M^\top$ and its diagonal elements are determined by the zero row-sum constraint $M\mathbf{1} = \mathbf{0}$. Since all restrictions imposed on M are linear, \mathcal{Q}_G is a linear subspace of $\mathbb{R}^{d \times d}$ of dimension m . The space \mathcal{Q}_G is thus isomorphic with \mathbb{R}^m : identify $M \in \mathcal{Q}_G$ with the elements M_{ij} for index pairs $(i, j) \in E$ such that $i < j$. Furthermore, the map π_G above, when restricted to \mathcal{Q}_G , is a linear isomorphism between \mathcal{Q}_G and \mathbb{R}^m ; in particular, it is continuous.

The set of precision matrices $\Theta \in \mathcal{P}_d^1$ with $\Theta_{ij} = 0$ for $(i, j) \notin \overline{E}$, i.e., those that figure in the matrix completion problem in Definition 4.1, is equal to

$$\mathcal{P}_G := \{M \in \mathcal{Q}_G : \ker M = \text{span}(\{\mathbf{1}\}), M \text{ positive semi-definite}\}.$$

Let $\lambda_1(S) \geq \dots \geq \lambda_d(S)$ be the d real eigenvalues of a symmetric matrix $S \in \mathbb{R}^{d \times d}$, counted with multiplicities and ordered decreasingly. We have

$$\mathcal{P}_G = \{M \in \mathcal{Q}_G : \lambda_{d-1}(M) > 0\}.$$

The functions λ_j are well-known to be Lipschitz on $\{S \in \mathbb{R}^{d \times d} : S = S^\top\}$. It follows that \mathcal{P}_G is open in \mathcal{Q}_G . Upon the above identification of \mathcal{Q}_G with \mathbb{R}^m via π_G , we can thus view \mathcal{P}_G as an open subset of \mathbb{R}^m . Formally, this subset is denoted as

$$\mathring{\mathcal{P}}_G := \pi_G(\mathcal{P}_G) \subset \mathbb{R}^m.$$

Let f_1 denote the inverse of the restriction of π_G to \mathcal{P}_G , that is,

$$\begin{aligned} f_1 : \mathring{\mathcal{P}}_G &\longrightarrow \mathcal{P}_G \\ v &\longmapsto M = \pi_G^{-1}(v) \end{aligned}$$

Since π_G was continuously invertible on the set \mathcal{Q}_G that contains \mathcal{P}_G , the function f_1 is a continuous bijection too.

The set of variogram matrices Γ corresponding to precision matrices M in \mathcal{P}_G is

$$\mathcal{D}_G := \theta^{-1}(\mathcal{P}_G) = \{\Gamma \in \mathcal{D}_d : \theta(\Gamma) \in \mathcal{P}_G\} = \{\theta^{-1}(M) : M \in \mathcal{P}_G\}.$$

By definition, these are exactly the variogram matrices $\Gamma \in \mathcal{D}_d$ of which the associated precision matrices $M = \theta(\Gamma)$ have zero elements $M_{ij} = 0$ for $(i, j) \notin \bar{E}$. In other words, \mathcal{D}_G corresponds to all possible solutions of the matrix completion problem in Definition 4.1 with respect to G .

By Proposition 4.5, a variogram matrix Γ in \mathcal{D}_G is uniquely determined by the values of Γ_{ij} for $(i, j) \in \bar{E}$. Since Γ has zero diagonal, this means that $\Gamma \in \mathcal{D}_G$ is uniquely determined by $\pi_G(\Gamma) = (\Gamma_{ij} : (i, j) \in E, i < j)$. Another way to say the same thing is that the map π_G restricted to \mathcal{D}_G is injective. For clarity, let

$$\mathring{\mathcal{D}}_G := \pi_G(\mathcal{D}_G) \subset \mathbb{R}^m$$

denote the image of \mathcal{D}_G under π_G and let f_2 denote the restriction of π_G to \mathcal{D}_G :

$$\begin{aligned} f_2 : \mathcal{D}_G &\longrightarrow \mathring{\mathcal{D}}_G \\ \Gamma &\longmapsto \pi_G(\Gamma). \end{aligned}$$

The inverse mapping of f_2 is the completion map \mathfrak{C}_G in Definition 4.6.² Knowing that f_2 is continuous (since π_G is continuous), we need to show that \mathfrak{C}_G is continuous as well.

To do this, consider the mapping

$$\begin{aligned} f : \mathring{\mathcal{P}}_G &\longrightarrow \mathring{\mathcal{D}}_G \\ v &\longmapsto (f_2 \circ \theta^{-1} \circ f_1)(v). \end{aligned}$$

The map f is represented schematically in Figure S.1. It sends the restriction $v = \pi_G(M) \in \mathring{\mathcal{P}}_G \subset \mathbb{R}^m$ of a precision matrix $M \in \mathcal{P}_G$ to the restriction $\pi_G(\Gamma) \in \mathring{\mathcal{D}}_G \subset \mathbb{R}^m$

²Actually, this comes with some abuse of notation: in Definition 4.6, the set $\mathring{\mathcal{D}}_G$ and the completion map \mathfrak{C}_G are defined with the placeholder “?” for pairs $(i, j) \notin \bar{E}$, whereas in this proof, these unknown elements are omitted altogether.

applied to M interpreted as a d^2 -dimensional vector. Then any $M \in B_\varepsilon$ satisfies

$$\begin{aligned}
\max_{v \in K} v^\top M v &= \max_{v \in K} \sum_i \sum_j M_{ij} v_i v_j \\
&\leq \max_{v \in K} \sum_i \sum_j \Gamma_{ij} v_i v_j + \varepsilon |v_i v_j| \\
&\leq \max_{v \in K} v^\top \Gamma v - \Delta_\Gamma / 2 \\
&= \Delta_\Gamma / 2 \\
&< 0.
\end{aligned}$$

Let $\mathcal{R}_d = \{M \in \mathbb{R}^{d \times d} : M = M^\top \wedge \text{diag}(M) = \mathbf{0}\}$ and observe that $\mathcal{R}_d \cap B_\varepsilon \subseteq \mathcal{D}_d$. For a set of matrices $S \subseteq \mathbb{R}^{d \times d}$ denote $S|_G = \{M|_G : M \in S\}$, with $\cdot|_G$ as in (4.4), and observe that

$$\mathcal{R}_d|_G \cap B_\varepsilon|_G \subseteq \mathcal{D}_d|_G = \mathring{\mathcal{D}}_d.$$

By assumption, any realization of $\hat{\Gamma}$ is always in $\mathcal{R}_d|_G$ and hence

$$\mathbb{P}(\hat{\Gamma} \in \mathring{\mathcal{D}}_d) = \mathbb{P}(\hat{\Gamma} \in B_\varepsilon|_G) = \mathbb{P}\left(\max_{(i,j) \in E} |\hat{\Gamma}_{ij} - \Gamma_{ij}| \leq \varepsilon\right) \rightarrow 1.$$

Together with the continuity from Lemma 5.2 this completes the proof. \square

S.6 Additional figures

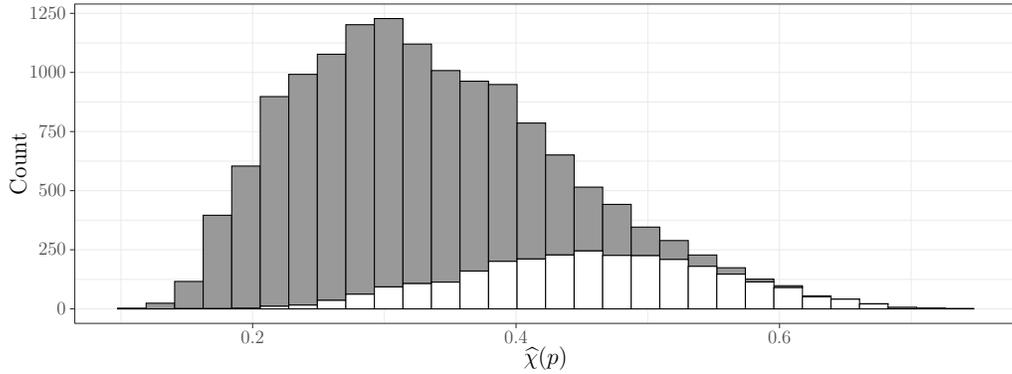


Figure S.2: Histogram of empirical $\hat{\chi}(p)$ for $p = 0.85$. Based on all pairs of airports. Values corresponding to two airports within the same cluster are shown in white, and values corresponding to two airports in two distinct clusters in gray.

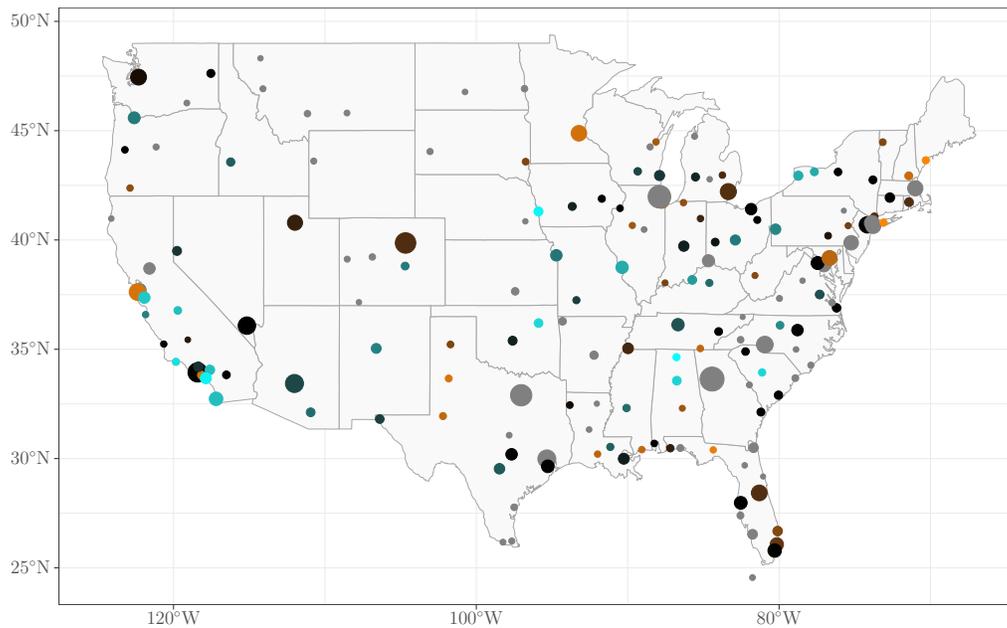


Figure S.3: Airports in the dataset. Color indicates the shape parameter of univariate generalized Pareto distributions, fitted for each airport using maximum likelihood estimation to the observations above the $p = 0.85$ quantile (black at zero, shade of cyan for negative, and orange for positive values). The size of each circle is proportional to the average number of daily flights at that airport.