# Spectral Sparsification of Laplacian-Constrained Gaussian and Hüsler–Reiss Graphical Models

**Ignacio Echave-Sustaeta Rodríguez** I.ECHAVE.SUSTAETA.RODRIGUEZ@TUE.NL
*Department of Mathematics and Computer Science, Eindhoven University of Technology*
*Eindhoven, The Netherlands*

**Aida Abiad** A.ABIAD@TUE.NL
*Department of Mathematics and Computer Science, Eindhoven University of Technology*
*Eindhoven, The Netherlands*

**Frank Röttger** F.ROTTGER@UTWENTE.NL
*Department of Applied Mathematics, University of Twente*
*Enschede, The Netherlands*

**Editors:** Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

## Abstract

Graph Laplacians encode graph structures in matrix form, and thus facilitate the application of linear algebra to graph theory. In statistics, two related families of probabilistic graphical models can be parameterized by graph Laplacians. The first one is the Laplacian-constrained Gaussian graphical model (LCGGM), which imposes that the (pseudo-)inverse covariance matrix of a Gaussian random vector is a Laplacian matrix. Applications include graph signal processing and network topology learning. The second one is the Hüsler–Reiss graphical model, which is considered as an extremal analog of the Gaussian graphical model, and can be used in extremal dependence modeling of floods, heatwaves, and financial losses. For both models, the restriction to positive edge weights in the graph Laplacian gives rise to an approach for graph structure learning that does not require tuning parameters. While these approaches yield a strong model fit in many settings, the resulting graph estimates are typically much denser than the underlying ground truth, limiting interpretability and scalability. In order to improve the accuracy of Laplacian-constrained graph learning, we propose to use *spectral graph sparsification* as a post-estimation operation. To do so, we replace the original Laplacian estimate by a sparser Laplacian that is spectrally close, and re-fit the model on the resulting graph. We refer to the two resulting methods as *Spectral-LCGGM* and *Spectral-HR*. We investigate the properties of the proposed estimators and show several theoretical results on their performance. Furthermore, we demonstrate that the newly proposed methods perform well: we run simulations for Spectral-LCGGM on Erdős–Rényi and stochastic block model graphs, and we showcase Spectral-HR on two real extremal data sets.

**Keywords:** Graph Laplacian, spectral sparsification, Laplacian-constrained Gaussian graphical models, graphical models for extremes, Hüsler–Reiss distribution, extremal $\mathrm{MTP}_2$.

## 1. Introduction

Probabilistic graphical models represent multivariate dependence through graphs, with edges connecting pairs of variables that interact directly after accounting for the others (Lauritzen, 1996). Two recent families of graphical models share the algebraic feature that their precision

© 2026 I. Echave-Sustaeta Rodríguez, A. Abiad & F. Röttger.

matrix is a connected weighted graph Laplacian. The first is the *Laplacian-constrained Gaussian graphical model* (LCGGM), in which that structure is imposed as a modeling constraint (Egilmez et al., 2017; Ying et al., 2020, 2021; Wang et al., 2022). LCGGMs are widely used in graph signal processing and network topology learning (Dong et al., 2016; Kalofolias, 2016; Egilmez et al., 2017), and are a degenerate-rank counterpart of Gaussian graphical models under multivariate total positivity of order 2 ($\mathrm{MTP}_2$) (Slawski and Hein, 2015; Lauritzen et al., 2019; Wang et al., 2020; Cai et al., 2023). Of these references, Slawski and Hein (2015) use related methodological ideas to the ones we employ here, since they also use the idea of refitting after processing the original estimator. The second is the *Hüsler–Reiss graphical model* (HRGM) under *extremal* $\mathrm{MTP}_2$ ($\mathrm{EMTP}_2$) (Engelke and Hitz, 2020; Röttger et al., 2023), where the Laplacian structure instead characterizes the positivity assumption.

To give some intuition, let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ be one simultaneous reading from a network of $d$ sensors, which is a signal on the nodes of an unknown graph. An LCGGM models $n$ such signals, after removing their network-wide average, as i.i.d. centered Gaussian vectors whose precision matrix $K$ is the weighted Laplacian of that graph. Up to constants its log-density is $-\frac{1}{2}x^\top Kx = -\frac{1}{2}\sum_{i<j} w_{ij}(x_i - x_j)^2$, so signals varying little across strongly connected nodes are the likely ones, and the zero pattern of $K$ is the Markov structure of the model. We are interested in recovering the underlying graph and the edge weights $w_{ij}$ from a sample of such signals.

On the other hand, Hüsler–Reiss graphical models instead address joint extremes. Take $x$ to be the daily discharges at $d$ gauging stations along the Danube, as in Section 4.2. Flood risk is governed by which stations are extreme simultaneously. Gaussian graphical models are unsuited here, since Gaussian vectors are asymptotically independent in the tail. Extreme value theory instead models the limit law of $x$ given that some component exceeds a high threshold, and the HRGM is the analog of the Gaussian graphical model for that limit (Engelke and Hitz, 2020): its parameter is a variogram matrix $\Gamma$, equivalently a precision matrix $\Theta$ whose zero pattern is an extremal conditional independence graph. Under $\mathrm{EMTP}_2$, a positive-dependence assumption natural for river networks and many other extremal applications (Röttger et al., 2023), $\Theta$ is again a connected graph Laplacian. We again would like to recover the graph and the edge weights from an appropriate sample of extreme events.

In both settings, natural estimators are maximum likelihood estimators (MLEs) under the Laplacian (positivity) constraint: the combinatorial graph Laplacian estimator (CGL) of Egilmez et al. (2017) for LCGGMs and the $\mathrm{EMTP}_2$ surrogate MLE of Röttger et al. (2023) for HRGMs. These estimators require no tuning parameter and fit the data well, but they return overly dense graphs. In fact, Röttger et al. (2023, Theorem 5.9) show that the latter asymptotically identifies a super-graph of the underlying ground truth. Beyond obscuring model identification, spurious edges slow down downstream graph-level computations. The standard remedy, $\ell_1$-penalization, does not help here: since the $\ell_1$-norm of the off-diagonal entries of a Laplacian equals its trace, the penalty shrinks all edge weights uniformly and does not promote sparsity, and in practice even yields denser graphs as the penalty grows (Ying et al., 2020). Penalized alternatives that do produce sparse graphs, for example the nonconvex formulations for LCGGMs of Ying et al. (2020) or the eglearn estimator for HRGMs (Engelke et al., 2026), require the choice of a tuning parameter and may incur a

measurable cost in fit quality. Our proposal also needs a tuning parameter, but it controls a spectral-approximation budget with an explicit Loewner certificate and edge bound rather than an opaque shrinkage level.

In this paper we employ techniques from spectral graph theory to maintain the fit quality of the unpenalized estimators while improving on their sparsity. Because the precision matrices of LCGGMs and HRGMs are already graph Laplacians, we can apply the deterministic linear-size sparsifier of Batson et al. (2012) (BSS), following the post-estimation strategy of Echave-Sustaeta Rodríguez et al. (2026) for the full-rank Gaussian $\mathrm{MTP}_2$ model. The recipe is the same in both settings: learn a dense graph estimate, sparsify the resulting precision matrix, and refit on the sparsified support. We call the resulting estimators *Spectral-LCGGM* and *Spectral-HR*. The fit survives this operation for a simple reason: in both models the log-likelihood depends on the precision matrix only through quadratic forms $z^\top K z$ and through $\log \mathrm{Det} K$. A spectral approximation with tolerance $\varepsilon$ preserves every quadratic form within a factor $1 \pm \varepsilon$ and changes $\log \mathrm{Det} K$ by between $(d-1)\log(1-\varepsilon)$ and $(d-1)\log(1+\varepsilon)$, whatever the underlying graph. Theorem 3.1 makes this precise. Recovery should improve too, although we argue this only heuristically: the sparsifier keeps just linearly many edges, and the ones it drops are those that are spectrally redundant, which under correct specification is what the dense MLE's spurious low-weight edges are. The refit then repairs the surviving weights. We do not, however, claim exact support recovery: a spectral sparsifier may in principle replace an edge by a spectrally equivalent one. Theoretically, we bound the change in log-likelihood induced by the sparsifier in both settings and establish model-class preservation, together with additional guarantees on variograms and the Hüsler–Reiss exponent measure density (Section 4). The analysis rests on a single identity on $\mathbf{1}^\perp$ combining a Bregman representation of the log-likelihood gap with the spectral approximation. Empirically, Spectral-LCGGM improves both graph recovery and test log-likelihood over the denser baselines at far fewer edges, and Spectral-HR retains the fit of the dense baselines on two extremal datasets at substantially reduced edge counts. Code for the experiments is available at `https://github.com/iechavesustaeta/SparsificationLaplacianGM`.

**Related work.** Spectral sparsification was introduced by Spielman and Teng (2004) and realized by various constructions (Spielman and Srivastava, 2011; Batson et al., 2012; Koutis et al., 2012). Closely related sparse approximate factorizations of Laplacians and symmetric diagonally dominant $M$-matrices arise from approximate Gaussian elimination (Kyng and Sachdeva, 2016). In statistics and machine learning, the barrier method behind the BSS sparsifier underlies the covariance-estimation bounds of Srivastava and Vershynin (2013), and sparsification has been used to accelerate sampling from Gaussian graphical models (Cheng et al., 2015), to precede Laplacian-regularized regression while preserving statistical validity (Sadhanala et al., 2016), and, in ridge form, to scale Laplacian-regularized semi-supervised learning (Calandriello et al., 2018). Relative to Echave-Sustaeta Rodríguez et al. (2026), whose post-estimation sparsifier and Bregman–Loewner analysis we follow, the LCGGM and $\mathrm{EMTP}_2$ Hüsler–Reiss settings require degenerate-rank versions of the bounds, that is Kullback–Leibler (KL) divergence on $\mathbf{1}^\perp$, pseudo-determinants, and trace-norm residuals through the projector $P$. Wang et al. (2022) instead learn a Laplacian-constrained graph by greedy spectral *densification*, adding the most spectrally critical edges to a sparse base graph, whereas we sparsify after estimation with explicit Loewner-order, log-likelihood and

KL guarantees. Beyond the estimators already named, other LCGGM solvers include those of Zhao et al. (2019) and Kumar et al. (2020).

## 2. Preliminaries

**Notation.** Throughout, $d$ is the number of variables, $[d] := \{1, \dots, d\}$, $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector, $\mathbf{1}^\perp := \{z \in \mathbb{R}^d : z^\top \mathbf{1} = 0\}$ its orthogonal complement, and $P := I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$ the orthogonal projector onto $\mathbf{1}^\perp$. For a symmetric matrix $A$, $A^+$ denotes the Moore–Penrose pseudoinverse, $\mathrm{Det}(A)$ the pseudo-determinant (product of the nonzero eigenvalues), $\mathrm{tr}(A)$ the trace, $\mathrm{diag}(A)$ the vector of diagonal entries, $\|A\|_{\mathrm{op}}$ the operator norm and $\|A\|_*$ the trace norm. If $A$ is positive semidefinite with kernel $\mathrm{span}(\mathbf{1})$, $A^{1/2}$ is its symmetric square root on $\mathbf{1}^\perp$ (and zero on $\mathrm{span}(\mathbf{1})$). $A \preceq B$ denotes the Loewner order (defined below). For a graph Laplacian $M$, $k(M)$ is the number of edges of its graph, i.e. the number of nonzero off-diagonal pairs. Estimators are marked as follows: a hat $(\widehat{K}, \widehat{\Theta})$ denotes a dense constrained maximum likelihood estimator, a tilde $(\widetilde{K}, \widetilde{\Theta})$ its spectral sparsifier, and a bar $(\overline{\Gamma}, \overline{\Sigma})$ an empirical summary statistic computed from the data. $S$ denotes a sample covariance matrix, and $T$ a generic matrix at which a log-likelihood is evaluated (Section 3.1). Proofs of all our results are collected in Appendix A.

### 2.1. Graphs, Laplacians, and Spectral Sparsification

We consider connected weighted graphs $G = (V, E, c)$ with symmetric weight function $c : V \times V \to [0, \infty)$. Their weighted Laplacian $L_G := \mathrm{Deg}(G) - A$, with $A = (c_{ij})$ the adjacency matrix and $\mathrm{Deg}(G) = \mathrm{diag}(\delta_1, \dots, \delta_{|V|})$, $\delta_i = \sum_j c_{ij}$, is positive semidefinite, satisfies $L_G\mathbf{1} = 0$ and has rank $|V| - 1$.

For symmetric $A, B \in \mathbb{R}^{|V|\times|V|}$ and a subspace $W \subseteq \mathbb{R}^{|V|}$ we write $A \preceq B$ *on* $W$ if $z^\top A z \le z^\top B z$ for every $z \in W$, and $A \succeq B$ symmetrically. This is the *Loewner order*, and the suffix is omitted when $W = \mathbb{R}^{|V|}$. For Laplacians the natural choice is $W = \mathbf{1}^\perp$, since their kernel contributes no information. For $\varepsilon \in (0, 1)$, a graph $G'$ on the same vertex set is an *$\varepsilon$-spectral approximation* of $G$ if

$$(1 - \varepsilon)\, L_G \;\preceq\; L_{G'} \;\preceq\; (1 + \varepsilon)\, L_G. \tag{1}$$

The construction of Batson et al. (2012), the BSS sparsifier, achieves (1) with linearly many edges.

**Theorem 2.1 (BSS sparsification)** *Let $L$ be a connected weighted Laplacian on $d$ vertices and fix $\eta > 1$. There is a deterministic polynomial-time algorithm that outputs a reweighted subgraph Laplacian $\widetilde{L}$ supported on at most $\lceil \eta(d-1) \rceil$ edges that satisfies*

$$L \preceq \widetilde{L} \preceq \kappa(\eta)\, L \qquad \text{on } \mathbf{1}^\perp, \qquad \kappa(\eta) := \left(\tfrac{\sqrt{\eta}+1}{\sqrt{\eta}-1}\right)^2.$$

*After rescaling $\widetilde{L}$ by the constant $2/(1 + \kappa(\eta))$, this becomes the symmetric $(1 \pm \varepsilon)$ Loewner bound $(1 - \varepsilon)L \preceq \widetilde{L} \preceq (1 + \varepsilon)L$ on $\mathbf{1}^\perp$ with $\varepsilon = (\kappa(\eta) - 1)/(\kappa(\eta) + 1) = 2\sqrt{\eta}/(\eta + 1)$.*

### 2.2. Laplacian-Constrained Gaussian Graphical Models

A *Laplacian-constrained Gaussian graphical model* (LCGGM) is a centered, degenerate Gaussian random vector $X$ with support $\mathbf{1}^\perp \subset \mathbb{R}^d$, whose precision matrix $K$ is a connected weighted graph Laplacian (Egilmez et al., 2017; Ying et al., 2020; Wang et al., 2022). Its density with respect to the Lebesgue measure on $\mathbf{1}^\perp$ is $f_K(x) = (2\pi)^{-(d-1)/2}\mathrm{Det}(K)^{1/2}e^{-x^\top Kx/2}$. For a sample covariance matrix $S \succeq 0$ obtained from $n$ i.i.d. copies of $X$, we write

$$\ell(K;S) := \log \mathrm{Det}(K) - \mathrm{tr}(KS), \tag{2}$$

so that the log-likelihood of the whole sample equals $\frac{n}{2}\,\ell(K;S)$ up to an additive constant. Writing

$$\mathcal{C} := \left\{K \in \mathbb{R}^{d\times d} : K = K^\top,\ K\mathbf{1} = 0,\ K_{ij} \le 0 \text{ for } i \ne j,\ \mathrm{rank}(K) = d-1\right\} \tag{3}$$

for the cone of connected graph Laplacians, this gives rise to the LCGGM-MLE given $S$,

$$\widehat{K} = \arg\max_{K\in\mathcal{C}}\ \ell(K;S). \tag{4}$$

LCGGMs are the Laplacian analog of the Gaussian $\mathrm{MTP}_2$ distributions (Lauritzen et al., 2019; Wang et al., 2020), adding zero row sums to the nonpositive off-diagonal constraint.

### 2.3. Hüsler–Reiss Graphical Models and graph Laplacians

Let $X = (X_1, \ldots, X_d)$ be a random vector with continuous marginal cumulative distribution functions $F_i$. After transforming to exponential margins $X_i^* := -\log(1 - F_i(X_i))$, the threshold-exceedance limit

$$\mathbb{P}(Y \le y) = \lim_{u\to\infty} \mathbb{P}\big(X^* - u\mathbf{1} \le y \mid \max_{i\in[d]} X_i^* \ge u\big) \tag{5}$$

yields, under multivariate regular variation, a multivariate (generalized) Pareto vector $Y$. Here $u$ is a high threshold and the conditioning event is that at least one coordinate exceeds it, so $Y$ describes $X$ given that an extreme event occurs somewhere. Assuming asymptotic dependence, $Y$ is supported on $\mathcal{E} := \{y \in \mathbb{R}^d : \max_i y_i > 0\}$ (Rootzén and Tajvidi, 2006; Hentschel et al., 2025).

A $d$-variate Hüsler–Reiss distribution (Hüsler and Reiss, 1989; Engelke and Hitz, 2020) is a multivariate Pareto distribution that can be parametrized by a conditionally negative definite variogram matrix

$$\Gamma \in \mathcal{D}^d := \left\{\Gamma \in \mathbb{R}^{d\times d} : \Gamma = \Gamma^\top,\ \mathrm{diag}(\Gamma) = 0,\ x^\top\Gamma x < 0\ \forall\, x \ne 0,\ x \perp \mathbf{1}\right\}.$$

The variogram entry $\Gamma_{ij}$ quantifies the strength of extremal dependence between $Y_i$ and $Y_j$: it equals the conditional variance $\mathrm{Var}(Y_i - Y_j \mid Y_k > 0)$, which does not depend on the conditioning coordinate $k$, and the bivariate extremal correlation is $\chi_{ij} = 2 - 2\Phi(\sqrt{\Gamma_{ij}}/2)$ with $\Phi$ the standard normal distribution function. Thus $\Gamma_{ij} \downarrow 0$ corresponds to complete extremal dependence and $\Gamma_{ij} \uparrow \infty$ to asymptotic independence (Engelke and Hitz, 2020).

The probability density of the Hüsler–Reiss distribution is obtained by normalizing a function $\lambda_\Gamma$, typically referred to as exponent measure density:

$$f_\Gamma(y) = \frac{\lambda_\Gamma(y)}{Z(\Gamma)}, \qquad \lambda_\Gamma(y) = \frac{1}{\sqrt{-\det \mathrm{CM}(2\pi\Gamma)}} \exp\Big(-\tfrac{1}{2}\begin{pmatrix} y & 1 \end{pmatrix} \mathrm{CM}(\Gamma)^{-1} \begin{pmatrix} y \\ 1 \end{pmatrix}\Big), \qquad y \in \mathcal{E}, \tag{6}$$

where $Z(\Gamma) := \int_{\mathcal{E}} \lambda_\Gamma(y)\,\mathrm{d}y$ denotes the normalizing constant, that is the exponent measure of $\mathcal{E}$. Here, $\mathrm{CM}(\Gamma) := \left(\begin{smallmatrix} -\Gamma/2 & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{smallmatrix}\right)$ is the Cayley–Menger matrix (Fiedler, 2011; Devriendt, 2022; Devriendt et al., 2026). Define $\Sigma := -\frac{1}{2}P\Gamma P$. The variogram is recovered from $\Sigma$ via $\Gamma = \mathrm{diag}(\Sigma)\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(\Sigma)^\top - 2\Sigma$. If $\Gamma \in \mathcal{D}^d$, then $\Sigma$ is positive semidefinite with kernel $\mathrm{span}(\mathbf{1})$ and has a Moore–Penrose pseudoinverse $\Theta = \Sigma^+$. In particular, $\Gamma_{ij} = \Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij} = (e_i - e_j)^\top \Theta^+ (e_i - e_j)$, which, when $\Theta$ is a graph Laplacian, is the effective resistance between nodes $i$ and $j$ in the graph with conductances $-\Theta_{ij}$ (Devriendt, 2022). Moreover, $\Theta_{ij} = 0$ is equivalent to the extremal conditional independence of $Y_i$ and $Y_j$ given the remaining coordinates (Hentschel et al., 2025). Furthermore, $\Theta$ being a graph Laplacian, that is $\Theta_{ij} \le 0$ for all $i \neq j$, is equivalent to an extremal notion of total positivity ($\mathrm{EMTP}_2$), compare Röttger et al. (2023).

For i.i.d. data in the domain of attraction of $Y$ as in (5), the empirical variogram $\overline{\Gamma}$ of Engelke and Volgushev (2022) is consistent for $\Gamma$ and serves as a summary statistic for estimation in HRGMs (Hentschel et al., 2025). To estimate $\Theta$ under $\mathrm{EMTP}_2$, Röttger et al. (2023) propose a surrogate MLE that maximizes the LCGGM log-likelihood at $\overline{\Sigma} = -\frac{1}{2} P \overline{\Gamma} P$, that is

$$\widehat{\Theta} = \arg\max_{\Theta \in \mathcal{C}} \ \ell(\Theta; \overline{\Sigma}), \tag{7}$$

with $\mathcal{C}$ as in (3).

## 3. Laplacian-Constrained Gaussian Graphical Models

### 3.1. Theory

The LCGGM-MLE $\widehat{K}$ obtained from a sample covariance $S$ via (4) is by construction a connected graph Laplacian, so Theorem 2.1 applies directly: there exists a sparsifier $\widetilde{K}$ that is also a connected graph Laplacian and satisfies the spectral approximation

$$(1-\varepsilon)\widehat{K} \preceq \widetilde{K} \preceq (1+\varepsilon)\widehat{K} \qquad \text{on } \mathbf{1}^\perp. \tag{8}$$

We write $D_{\mathrm{KL}}^{\perp}(\widehat{K} \,\|\, \widetilde{K}) := \frac{1}{2}[\mathrm{tr}(\widetilde{K}\widehat{K}^+) - (d-1) - \log \mathrm{Det}(\widetilde{K}\widehat{K}^+)]$ for the KL divergence between the two centered degenerate Gaussians on $\mathbf{1}^\perp$ with precisions $\widehat{K}$ and $\widetilde{K}$, derived in Lemma A.3.

Our goal is to control the LCGGM log-likelihood gap $\ell(\widetilde{K}; T) - \ell(\widehat{K}; T)$ induced by replacing $\widehat{K}$ by its sparsifier $\widetilde{K}$. The sample covariance $S$ of the training data is the input from which $\widehat{K}$ is computed. The matrix $T$ is the one at which the two fitted models are compared: $T = S$ when the fit is assessed on the training data (in-sample), and $T = S_{\mathrm{test}}$, the sample covariance of held-out data, when it is assessed on new data. In the Hüsler–Reiss setting of Section 4, the empirical summary $\overline{\Sigma}$ built from the variogram takes the role of $S$. Since $K\mathbf{1} = 0$ for every Laplacian $K$, we have $\mathrm{tr}(KT) = \mathrm{tr}(KPTP)$ and hence

$\ell(K;T) = \ell(K;PTP)$: the log-likelihood only depends on the projection $PTP$ of $T$ onto $\mathbf{1}^\perp$. We may and do therefore assume $T\mathbf{1} = 0$ throughout. This holds automatically for the sample covariance of data supported on $\mathbf{1}^\perp$, as in the LCGGM, and by construction for $\overline{\Sigma} = -\frac{1}{2}P\overline{\Gamma}P$.

**Theorem 3.1** *Let $\varepsilon \in (0,1)$, let $\widehat{K}, \widetilde{K}$ be connected graph Laplacians satisfying (8), and let $T \succeq 0$ be symmetric with $T\mathbf{1} = 0$. Then*

$$\ell(\widetilde{K};T) - \ell(\widehat{K};T) = -2\, D_{\mathrm{KL}}^{\perp}(\widehat{K} \,\|\, \widetilde{K}) - \mathrm{tr}\big((\widetilde{K} - \widehat{K})(T - \widehat{K}^{+})\big), \tag{9}$$

*and, setting $R_{\widehat{K}}(T) := \big\|\widehat{K}^{1/2}(T - \widehat{K}^{+})\widehat{K}^{1/2}\big\|_{*}$,*

$$-\frac{(d-1)\varepsilon^2}{2(1-\varepsilon)} - \varepsilon\, R_{\widehat{K}}(T) \;\le\; \ell(\widetilde{K};T) - \ell(\widehat{K};T) \;\le\; \varepsilon\, R_{\widehat{K}}(T). \tag{10}$$

The identity (9) splits the gap into a structural cost, the KL term, which is the price of moving away from $\widehat{K}$, and a data residual that vanishes at $T = \widehat{K}^{+}$. The quantity $R_{\widehat{K}}(T)$ measures that residual in the geometry induced by $\widehat{K}$. Any $\widetilde{K}$ satisfying (8) is automatically a valid LCGGM precision matrix, that is an element of $\mathcal{C}$, with algebraic connectivity (its smallest nonzero eigenvalue) preserved within $1 \pm \varepsilon$. Consequently the gap is non-positive at $T = S$, since $\widehat{K}$ maximizes $\ell(\cdot\,; S)$ over $\mathcal{C}$ and $\widetilde{K}$ is feasible for that program. At $T = S_{\mathrm{test}}$ the residual can take either sign, so the sparsifier may improve the held-out log-likelihood.

Theorem 3.1 guarantees that any spectral sparsifier of $\widehat{K}$ loses only an $O(\varepsilon)$ amount of in-sample log-likelihood, which motivates the three-step pipeline of Algorithm 1. The same algorithm applies to the Hüsler–Reiss setting of Section 4, where the surrogate HR-MLE (7) is the same convex program with $\overline{\Sigma}$ in place of $S$, so Algorithm 1 is stated with a generic input matrix.

**Algorithm 1:** Spectral-LCGGM and Spectral-HR (unified).

**Input:** Input matrix $S \in \mathbb{R}^{d\times d}$. BSS parameter $\eta > 1$.
**Output:** Sparse precision Laplacian $\widetilde{K}^{\mathrm{refit}}$.
**Step 1: Dense MLE:** $\widehat{K} \leftarrow \arg\max\{\ell(K;S) : K \in \mathcal{C}\}$, with $\mathcal{C}$ the cone (3)
**Step 2: BSS sparsification:** run Theorem 2.1 on $\widehat{K}$ at $\eta$, obtaining $\widetilde{K}$ and its support $\widetilde{E} \leftarrow \{(i,j) : \widetilde{K}_{ij} \neq 0,\, i \neq j\}$. Only $\widetilde{E}$ is used below, so the rescaling of Theorem 2.1 need not be carried out in practice
**Step 3: Refit: return** $\widetilde{K}^{\mathrm{refit}} \leftarrow \arg\max\{\ell(K;S) : K \in \mathcal{C},\, \mathrm{supp}(K) \subseteq \widetilde{E}\}$

The guarantees of Theorem 3.1, stated for the sparsifier $\widetilde{K}$ of Step 2, transfer to the estimator that Algorithm 1 actually returns. Step 3 maximizes the same objective over a feasible set containing $\widetilde{K}$, so $\ell(\widetilde{K}^{\mathrm{refit}};S) \ge \ell(\widetilde{K};S)$ and the lower bound of (10) at $T = S$ carries over to $\widetilde{K}^{\mathrm{refit}}$. Corollary A.4 in Appendix A.2 states this, for Spectral-HR as well.

Tuning proceeds through $\eta$, selected by BIC on the training data, with smaller $\eta$ giving sparser subgraphs at looser spectral certificates. At the selected $\eta$ the Loewner certificate is loose, so Theorem 3.1 and Corollary A.4 should be read as qualitative guarantees that sparsification is safe rather than as numerically binding bounds.

Regarding the computational cost, Step 1 is a convex program, solvable to global optimality by standard first- and second-order methods (Egilmez et al., 2017; Ying et al., 2020; Zhao et al., 2019) at $O(d^3)$ per iteration, Step 2 costs $O(\eta\, d^3 m)$ with $m$ the number of edges of $\widehat{K}$, and Step 3 is Step 1 restricted to at most $\lceil \eta(d-1) \rceil$ edges. Since Step 1 does not depend on $\eta$, the whole $\eta$-path costs one dense fit plus one BSS run and one sparse refit per grid point, and Appendix B gives our solver and the constants behind these rates, with running times in Sections 3.2 and 4.2.

### 3.2. Experiments

We evaluate Spectral-LCGGM on data simulated from an LCGGM with ground-truth precision $K^\star$ and edge set $E^\star$, both specified below. For each method, we report the edge count $|\widehat{E}|$ (off-diagonals exceeding $10^{-3}$ in magnitude), the precision $|\widehat{E} \cap E^\star|/|\widehat{E}|$, the recall $|\widehat{E} \cap E^\star|/|E^\star|$, their harmonic mean F1, and the LCGGM test log-likelihood $\frac{n_{\text{test}}}{2}\, \ell(\widehat{K}; S_{\text{test}})$ evaluated on a held-out sample of size $n_{\text{test}}$. Tuning parameters are selected on the training split by the Bayesian information criterion $\mathrm{BIC}(M) := -n_{\text{train}}\, \ell(M; S_{\text{train}}) + k(M)\, \log n_{\text{train}}$, with $\ell$ the LCGGM log-likelihood (2) and $S_{\text{train}}$ the training covariance.

We compare Spectral-LCGGM (Algorithm 1) against two baselines that span the sparsity–fit trade-off: CGL (Egilmez et al., 2017), the unpenalized LCGGM-MLE, and NGL-SCAD (Ying et al., 2020), a nonconvex SCAD-penalized LCGGM solver. The Graphical Lasso is excluded because $\ell_1$ penalization does not sparsify Laplacians (Section 1).

The ground-truth precision matrix $K^\star$ is the Laplacian of a weighted random graph, with edge weights drawn independently from the symmetric triangular distribution on $[0.5, 1.5]$ (the average of two independent $\mathrm{Unif}[0.5, 1.5]$ draws) to avoid eigenvalue multiplicities. We consider two random graph families. In an Erdős–Rényi (ER) graph $G(d, p)$ (Erdős and Rényi, 1959; Gilbert, 1959) each of the $\binom{d}{2}$ possible edges is present independently with probability $p$. We take $p = 4/d$, so the expected degree is $4(d-1)/d \approx 4$ regardless of $d$. In a stochastic block model (SBM) (Holland et al., 1983) the nodes are partitioned into communities and edges are present independently with probability $p_{\text{in}}$ within a community and $p_{\text{out}}$ between communities. We use five equal-sized blocks with $p_{\text{in}} = 0.12$ and $p_{\text{out}} = 0.01$, inducing community structure. In both cases the graph is resampled until it is connected. Each topology is realized at two dimensions, $d \in \{100, 200\}$, with $n_{\text{train}} = 4d$ training and $n_{\text{test}} = 4{,}000$ test observations, over $B = 10$ replications. Within each setting the graph and its weights are drawn once and held fixed, so $K^\star$ and $|E^\star|$ are common to all $B$ replications and only the data are resampled. The standard errors below therefore quantify sampling variability conditional on that ground-truth graph, not across draws of the graph.

Table 1 reports the BIC-selected results. Across all four settings the unpenalized LCGGM-MLE (CGL) recovers a graph much denser than the truth, with high recall and very low precision. NGL-SCAD (BIC-selected $\alpha$) brings the edge count closer to the truth with substantially improved F1. Spectral-LCGGM consistently delivers edge counts within 7% of the truth and achieves the highest F1 ($\geq 0.93$) in every setting, while also obtaining the best test log-likelihood. The test log-likelihood gap to NGL-SCAD is narrow, reflecting that NGL-SCAD also fits the data well, so Spectral-LCGGM's advantage is primarily on graph recovery. Spectral-LCGGM is also the most expensive method here, ten to thirteen times the cost of the CGL fit alone and, at $d = 200$ with community structure, some sixty times

that of NGL-SCAD. Most of this is the tuning sweep rather than the method itself, and the cost of a grid point grows with $\eta$ because more edges are retained and the refit is correspondingly larger. CGL and Spectral-LCGGM use our own R coordinate-descent solver at a tight tolerance, whereas NGL-SCAD is the `sparseGraph` package, so timings are for illustrative purposes.

| Setting | Method | Edges | Precision | Recall | F1 | Test log-lik. | Time (s) |
|---|---|---|---|---|---|---|---|
| ER, $d = 100$ $\lvert E^\star\rvert = 176$ | CGL | 503.3 (3.6) | 0.350 (0.003) | **1.000** (0.000) | 0.518 (0.003) | −8,155 (148) | 43 (2) |
| | NGL-SCAD | 243.7 (2.4) | 0.722 (0.007) | 0.998 (0.001) | 0.838 (0.005) | −7,215 (151) | 32 (3) |
| | Spectral-LCGGM | **187.2** (1.4) | **0.936** (0.007) | 0.995 (0.001) | **0.965** (0.003) | **−6,771** (157) | 510 (4) |
| SBM, $d = 100$ $\lvert E^\star\rvert = 171$ | CGL | 449.5 (3.0) | 0.380 (0.003) | **0.999** (0.001) | 0.551 (0.003) | −27,277 (126) | 48 (3) |
| | NGL-SCAD | 232.8 (2.3) | 0.734 (0.008) | 0.998 (0.001) | 0.846 (0.005) | −26,364 (132) | 35 (2) |
| | Spectral-LCGGM | **179.8** (1.5) | **0.945** (0.007) | 0.994 (0.002) | **0.969** (0.003) | **−26,090** (155) | 495 (5) |
| ER, $d = 200$ $\lvert E^\star\rvert = 391$ | CGL | 1,257.7 (5.7) | 0.311 (0.001) | **1.000** (0.000) | 0.474 (0.002) | 19,483 (282) | 817 (17) |
| | NGL-SCAD | 496.7 (8.0) | 0.787 (0.012) | 0.998 (0.001) | 0.880 (0.008) | 21,037 (261) | 282 (14) |
| | Spectral-LCGGM | **418.2** (2.1) | **0.935** (0.005) | 0.999 (0.000) | **0.966** (0.002) | **21,445** (267) | 9,251 (58) |
| SBM, $d = 200$ $\lvert E^\star\rvert = 623$ | CGL | 1,882.9 (12.8) | 0.330 (0.002) | **0.997** (0.001) | 0.496 (0.002) | 248,229 (235) | 1,002 (38) |
| | NGL-SCAD | 778.9 (3.3) | 0.769 (0.004) | 0.961 (0.002) | 0.854 (0.003) | 249,516 (221) | 216 (7) |
| | Spectral-LCGGM | **643.4** (2.8) | **0.918** (0.004) | 0.948 (0.002) | **0.933** (0.003) | **250,022** (215) | 12,701 (94) |

Table 1: LCGGM simulation, settings as in the text. Means over $B = 10$ replications with standard errors in parentheses. Time is wall-clock seconds per replication for the complete procedure, including tuning (eight $\alpha$ values for NGL-SCAD, fourteen $\eta$ values on $[1.5, 32]$ for Spectral-LCGGM, none for CGL). The BIC-selected $\eta$ is interior to that grid in every replication. Bold: best per metric within each $(d, \text{topology})$ block, closest to $\lvert E^\star\rvert$ for edges.

## 4. Hüsler–Reiss Graphical Models

### 4.1. Theory

Under $\text{EMTP}_2$, the precision matrix $\Theta$ is a graph Laplacian (Section 2.3), so Theorem 2.1 applies directly to $\Theta$. Moreover, by (7) the surrogate Hüsler–Reiss MLE $\widehat{\Theta}$ is the LCGGM-MLE (4) evaluated at $\overline{\Sigma}$ in place of the sample covariance. No separate analysis is therefore required for the Hüsler–Reiss case: the LCGGM theory of Section 3.1 works immediately under the substitution $\widehat{K} \to \widehat{\Theta}$, $\widetilde{K} \to \widetilde{\Theta}$, $S \to \overline{\Sigma}$, $\widehat{K}^+ \to \widehat{\Sigma} = \widehat{\Theta}^+$, with the spectral approximation

$$(1 - \varepsilon)\widehat{\Theta} \preceq \widetilde{\Theta} \preceq (1 + \varepsilon)\widehat{\Theta} \qquad \text{on } \mathbf{1}^\perp \tag{11}$$

in place of (8). The unified estimator of Algorithm 1, run with $\widehat{\Theta}$ and input matrix $\overline{\Sigma}$, produces Spectral-HR. Concretely, the Bregman identity (9), the bound (10) and Corollary A.4 hold verbatim for the pair $(\widehat{\Theta}, \widetilde{\Theta})$ and any symmetric $T \succeq 0$ with $T\mathbf{1} = 0$. Only the residual acquires a new reading: evaluating at the data summary $T = \overline{\Sigma}$ gives $R_{\widehat{\Theta}}(\overline{\Sigma}) = \big\|\widehat{\Theta}^{1/2}(\overline{\Sigma} - \widehat{\Sigma})\widehat{\Theta}^{1/2}\big\|_* = \frac{1}{2}\big\|\widehat{\Theta}^{1/2} P(\overline{\Gamma} - \widehat{\Gamma}) P\,\widehat{\Theta}^{1/2}\big\|_*$, where $\widehat{\Gamma}$ is the variogram of $\widehat{\Theta}$. The second form shows that the residual measures the discrepancy between the empirical variogram $\overline{\Gamma}$ and the fitted variogram $\widehat{\Gamma}$. If the data-generating Hüsler–Reiss distribution is $\text{EMTP}_2$ and the variogram estimator is consistent, as is the empirical variogram in the

domain-of-attraction setting of Engelke and Volgushev (2022, Theorem 1) provided the number of exceedances grows at the rate required there, then $\widehat{\Theta}$ is consistent for $\Theta$ (Röttger et al., 2023, Theorem 5.7), hence $\overline{\Sigma} - \widehat{\Sigma} \to 0$ and $R_{\widehat{\Theta}}(\overline{\Sigma}) \to 0$ in probability, so the residual term in (10) vanishes asymptotically. Any sparsifier $\widetilde{\Theta}$ satisfying (11) is again a connected graph Laplacian, hence the precision of an $\mathrm{EMTP}_2$ Hüsler–Reiss distribution.

The two guarantees below are stated for an arbitrary pair of connected graph Laplacians related by a spectral approximation, so by (11) they apply to the estimated pair, with $\widehat{\Theta}$ in place of $\Theta$. Since $\Gamma_{ij} = (e_i - e_j)^\top \Theta^+ (e_i - e_j)$ is the effective resistance between $i$ and $j$ (Section 2.3), Theorem 4.1 says that a spectral sparsifier preserves all effective resistances, which is the reading in the context of extremes of the fact underlying the sampling scheme of Spielman and Srivastava (2011).

**Theorem 4.1** *Let $\varepsilon \in (0,1)$ and let $\Theta, \widetilde{\Theta} \in \mathcal{C}$ satisfy $(1-\varepsilon)\Theta \preceq \widetilde{\Theta} \preceq (1+\varepsilon)\Theta$ on $\mathbf{1}^\perp$, with associated variograms $\Gamma, \widetilde{\Gamma}$. For every pair $i \neq j$,*

$$\frac{\Gamma_{ij}}{1+\varepsilon} \;\le\; \widetilde{\Gamma}_{ij} \;\le\; \frac{\Gamma_{ij}}{1-\varepsilon}. \tag{12}$$

*In other words, each entry changes by at most a multiplicative factor, $\widetilde{\Gamma}_{ij}/\Gamma_{ij} \in [1/(1+\varepsilon),\, 1/(1-\varepsilon)]$, uniformly over all pairs and independently of $d$ and of the smallest nonzero eigenvalue of $\Theta$.*

The multiplicative variogram control transfers, via the Cayley–Menger representation (6), to a pointwise bound on the Hüsler–Reiss logarithmic exponent measure density. Before stating it, define

$$M_\Theta \;:=\; \big\|\mathrm{CM}(\Gamma)^{-1}\big\|_{\mathrm{op}} \;\le\; \|\Theta\|_{\mathrm{op}} + \|q\|_2 + |\sigma^2|, \tag{13}$$

where $\Theta \in \mathbb{R}^{d\times d}$, $q \in \mathbb{R}^d$, and $\sigma^2 \in \mathbb{R}$ are the blocks of $\mathrm{CM}(\Gamma)^{-1} = \left(\begin{smallmatrix} \Theta & q \\ q^\top & \sigma^2 \end{smallmatrix}\right)$, with $\Theta = \Sigma^+$ in the top-left $d \times d$ block (Fiedler–Bapat identity, see Fiedler, 2011; Bapat, 2014; Devriendt, 2022). Set also

$$\tau_\varepsilon \;:=\; \frac{\varepsilon\, \|\Gamma\|_{\mathrm{op}}}{2(1-\varepsilon)}. \tag{14}$$

**Corollary 4.2** *Let $\varepsilon \in (0,1)$ and let $\Theta, \widetilde{\Theta} \in \mathcal{C}$ satisfy $(1-\varepsilon)\Theta \preceq \widetilde{\Theta} \preceq (1+\varepsilon)\Theta$ on $\mathbf{1}^\perp$, with associated variograms $\Gamma, \widetilde{\Gamma}$, and suppose $\tau_\varepsilon M_\Theta < 1$, with $M_\Theta$ as in (13) and $\tau_\varepsilon$ as in (14). Then, for the exponent measure density $\lambda_\Gamma$ of (6) and every $y \in \mathcal{E}$,*

$$\big|\log \lambda_{\widetilde{\Gamma}}(y) - \log \lambda_\Gamma(y)\big| \;\le\; \frac{d+1}{2} \cdot \frac{\tau_\varepsilon M_\Theta}{1 - \tau_\varepsilon M_\Theta} \;+\; \frac{\|y\|_2^2 + 1}{2} \cdot \frac{\tau_\varepsilon M_\Theta^2}{1 - \tau_\varepsilon M_\Theta}. \tag{15}$$

*Averaging (15) over a sample $Y_1, \dots, Y_n \in \mathcal{E}$ gives the same bound for $n^{-1}\big|\sum_k [\log \lambda_{\widetilde{\Gamma}}(Y_k) - \log \lambda_\Gamma(Y_k)]\big|$, with $\|y\|_2^2$ replaced by $n^{-1}\sum_k \|Y_k\|_2^2$.*

Both bounds are $O(\varepsilon)$ as $\varepsilon \downarrow 0$. Unlike Theorem 4.1, however, they are not free of the spectrum of $\Theta$: they depend on its conditioning through $M_\Theta$ and $\|\Gamma\|_{\mathrm{op}}$. The sample-average form controls the data-dependent part of $\ell^{\mathrm{HR}}(\Theta; Y_{1:n}) = \sum_k \log \lambda_\Gamma(Y_k) - n \log Z(\Gamma)$, leaving the deterministic term $\log Z(\widetilde{\Gamma}) - \log Z(\Gamma)$, for which we do not develop an explicit bound.

The hypothesis $\tau_\varepsilon M_\Theta < 1$ holds for any fixed $\Theta$ once $\varepsilon$ is small enough, but it is never automatically satisfied: Remark A.5 in Appendix A.3 shows that it forces $\varepsilon < 1/2$, that is $\eta > 7 + 4\sqrt{3} \approx 13.9$.

### 4.2. Experiments

We evaluate Spectral-HR against the EMTP$_2$-MLE (Röttger et al., 2023) and the two majority-vote variants of eglearn (Engelke et al., 2026), neighborhood-selection (eglearn-NS) and graphical-lasso (eglearn-glasso), on two Hüsler–Reiss real-data benchmark datasets. The upper Danube river-network discharge dataset (Asadi et al., 2015; Engelke and Hitz, 2020) is too small to split once the data are thresholded, so methods are compared by BIC and AIC on the full thresholded sample. The Texas Cluster of U.S. airports (Hentschel et al., 2025) is much larger, so we use the standard train/test date split of Hentschel et al. (2025) and held-out Hüsler–Reiss log-likelihood as the headline metric. In both experiments the log-likelihood used for selection and held-out evaluation is the actual Hüsler–Reiss log-likelihood $\ell^{\mathrm{HR}}(\widehat{\Theta}) := \sum_k \log f_{\widehat{\Gamma}}(Y_k)$.

**Upper Danube river-network discharges (Asadi et al., 2015; Engelke and Hitz, 2020).** The data are daily seasonal-residual discharges at $d = 31$ gauging stations along the upper Danube and its tributaries, with $n = 428$ exceedance-cluster days as bundled in the `graphicalExtremes` R package (Engelke et al., 2024). Since there is no canonical extremal ground-truth graph for river discharges, we compare methods purely on in-sample model selection. At threshold $p = 0.90$ the thresholded sample has $n^{\mathrm{thr}} = 117$ exceedances, on which we evaluate both

$$\mathrm{BIC}(\widehat{\Theta}) \;=\; -2\,\ell^{\mathrm{HR}}(\widehat{\Theta}) \;+\; k(\widehat{\Theta})\,\log n^{\mathrm{thr}}, \qquad \mathrm{AIC}(\widehat{\Theta}) \;=\; -2\,\ell^{\mathrm{HR}}(\widehat{\Theta}) \;+\; 2\,k(\widehat{\Theta}), \tag{16}$$

where $k(\widehat{\Theta})$ is the edge count of the graph given by $\widehat{\Theta}$. Each tunable method is selected twice: once by BIC, once by AIC. The EMTP$_2$-MLE has no tuning parameter. Tuning grids are $\rho \in [0.005, 0.5]$ for eglearn-NS, $\rho \in [0.005, 0.75]$ (20 equidistant points) for eglearn-glasso, and $\eta \in [1.1, 100]$ for Spectral-HR.

Table 2 reports the BIC- and AIC-selected fits. Spectral-HR gives the sparsest fits by a clear margin (46–53 edges, vs 99–130 for eglearn-NS, 171 for eglearn-glasso, and 67 for EMTP$_2$). eglearn-NS attains both the best BIC and the best AIC, at much higher edge counts than Spectral-HR. Figure 1 in Appendix C shows the river flow graph together with the EMTP$_2$ and Spectral-HR estimates. Against the dense EMTP$_2$ baseline, Spectral-HR improves both criteria ($\Delta$BIC $= -71.5$ at 46 edges, $\Delta$AIC $= -28.6$ at 53 edges, that is 69% and 79% of its edge count respectively). It trails the eglearn-NS winners on the absolute criteria (88 BIC points / 302 AIC points behind), trading information-criterion fit for substantially sparser graphs.

Spectral-HR's edge count grows monotonically over $\eta$ and stays below EMTP$_2$'s 67 edges across the entire grid, whereas eglearn-NS admits a usable fit only for $\rho < 0.132$ and eglearn-glasso attains its joint BIC/AIC minimum at the second grid point ($\rho = 0.044$, 171 edges). The full sweeps, in Table 3 (Appendix D), show the eglearn fits decaying for stronger sparsity settings, with both eglearn-NS criteria rising above the dense EMTP$_2$ baseline past $\rho = 0.05$. They also permit a comparison at matched sparsity, which is the fairest reading of Table 2: at a common edge count the complexity penalties cancel, so the two methods differ only through the log-likelihood. At 56 edges eglearn-NS ($\rho = 0.091$) attains BIC 1,519.4 and AIC 1,364.8, against 1,281.8 and 1,127.2 for Spectral-HR ($\eta = 50$), better by 237.6 points on both criteria, that is 118.8 nats of Hüsler–Reiss log-likelihood. eglearn-NS wins the unconstrained comparison of Table 2 only by spending more than twice as many edges.

| Method | Selected by | Param | Edges | BIC | AIC | Time (s) |
|---|---|---|---|---|---|---|
| $\mathrm{EMTP}_2$ | (none) | – | 67 | 1,336.02 | 1,150.96 | 1.0 |
| eglearn-NS | BIC | 0.032 | 99 | **1,176.95** | 903.49 | 9.1 |
| | AIC | 0.023 | 130 | 1,179.38 | **820.29** | |
| eglearn-glasso | BIC & AIC | 0.044 | 171 | 1,342.75 | 870.42 | 15.2 |
| Spectral-HR | BIC | 4 | **46** | 1,264.51 | 1,137.45 | 8.2 |
| | AIC | 25 | 53 | 1,268.74 | 1,122.34 | |

Table 2: Upper Danube river-network discharges. Whole dataset used for fitting and for both selection criteria (16). Each tunable method appears once per criterion, with rows merged when both criteria agree, as for eglearn-glasso. Time is wall-clock seconds for each method's complete sweep (17 values of $\rho$ for eglearn-NS, 20 for eglearn-glasso, 20 values of $\eta$ for Spectral-HR), shared by both selections and comparable within a table only. Bold: lowest BIC, lowest AIC and smallest edge count across all rows.

**Texas Cluster of U.S. airports (Hentschel et al., 2025).** We also evaluate our method against the same baselines on this standard extremal dataset. Spectral-HR gives the sparsest fit under both criteria, and under AIC it comes within 34 nats of the dense $\mathrm{EMTP}_2$ baseline on held-out data at 85% of its edges. The full setup and results are in Appendix E.

## 5. Discussion

We extended the post-estimation spectral-sparsification approach of Echave-Sustaeta Rodríguez et al. (2026) from full-rank Gaussian $\mathrm{MTP}_2$ models to the LCGGM and the $\mathrm{EMTP}_2$ HRGM, whose precision matrices are connected graph Laplacians. Both are fitted by the same convex program, so one algorithm serves both, and a single Bregman–Loewner identity on $\mathbf{1}^{\perp}$ carries the analysis. The sparsified matrices remain valid parameters of their model class, with every Hüsler–Reiss variogram entry preserved multiplicatively over all pairs. Empirically, Spectral-LCGGM attains the best F1 and held-out log-likelihood in all four settings, at a computational price, and Spectral-HR gives the sparsest fits on both datasets, improving on the dense $\mathrm{EMTP}_2$ baseline on the Danube, where eglearn-NS attains better criteria only at twice the edges.

Finally, several research directions remain open. The refit could impose structure beyond positivity, and the almost-linear-time constructions of Lee and Sun (2018) could replace BSS in Step 2, whose cost grows with $\eta$. On the theoretical side, sharper or data-dependent guarantees would be valuable to improve our first analyses of the problem. The most interesting direction is structure learning: whether support recovery is achievable for either estimator, which would also call for simulation studies in the HRGM setting.

## Acknowledgements

The research of Aida Abiad is supported by NWO (Dutch Research Council) through the grant VI.Vidi.213.085. We would like to thank the anonymous reviewers and Program Chairs for their helpful comments and suggestions that helped improve the quality of the manuscript.

## References

Peiman Asadi, Anthony C. Davison, and Sebastian Engelke. Extremes on river networks. *The Annals of Applied Statistics*, 9(4):2023–2050, 2015. doi: 10.1214/15-AOAS863.

Ravindra B. Bapat. *Graphs and Matrices*. Universitext. Springer, London, 2nd edition, 2014. doi: 10.1007/978-1-4471-6569-9.

Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012. doi: 10.1137/090772873.

Jian-Feng Cai, José Vinícius de Miranda Cardoso, Daniel P. Palomar, and Jiaxi Ying. Fast projected Newton-like method for precision matrix estimation under total positivity. In *Advances in Neural Information Processing Systems*, volume 36, pages 73348–73370. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e878c8f38381d0964677fb9536c494ee-Paper-Conference.pdf.

Daniele Calandriello, Alessandro Lazaric, Ioannis Koutis, and Michal Valko. Improved large-scale graph learning through ridge spectral sparsification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 688–697. PMLR, 2018. URL https://proceedings.mlr.press/v80/calandriello18a.html.

Dehua Cheng, Yu Cheng, Yan Liu, Richard Peng, and Shang-Hua Teng. Efficient sampling for Gaussian graphical models via spectral sparsification. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 364–390. PMLR, 2015. URL http://proceedings.mlr.press/v40/Cheng15.html.

Karel Devriendt. *Graph geometry from effective resistances*. PhD thesis, University of Oxford, 2022.

Karel Devriendt, Ignacio Echave-Sustaeta Rodríguez, and Frank Röttger. Extremal conditional independence for Hüsler–Reiss distributions via modular functions. *arXiv preprint arXiv:2601.21931*, 2026.

Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst. Learning Laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23):6160–6173, 2016. doi: 10.1109/TSP.2016.2602809.

Ignacio Echave-Sustaeta Rodríguez, Aida Abiad, and Frank Röttger. Learning Gaussian graphical models under total positivity via spectral graph sparsification, 2026. URL https://arxiv.org/abs/2605.17154.

Hilmi E. Egilmez, Eduardo Pavez, and Antonio Ortega. Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, 2017. doi: 10.1109/JSTSP.2017.2726975.

Sebastian Engelke and Adrien S. Hitz. Graphical models for extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):871–932, 2020. doi: 10.1111/rssb.12355.

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

Sebastian Engelke, Adrien S. Hitz, Nicola Gnecco, and Manuel Hentschel. *graphicalExtremes: Statistical Methodology for Graphical Extreme Value Models*, 2024. URL https://CRAN.R-project.org/package=graphicalExtremes. R package version 0.3.3.

Sebastian Engelke, Michaël Lalancette, and Stanislav Volgushev. Learning extremal graphical structures in high dimensions. *The Annals of Statistics*, 54(3):1205–1231, 2026. doi: 10.1214/25-AOS2594.

Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

Miroslav Fiedler. *Matrices and Graphs in Geometry*, volume 139 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2011. doi: 10.1017/CBO9780511973611.

Edgar N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959. doi: 10.1214/aoms/1177706098.

Manuel Hentschel, Sebastian Engelke, and Johan Segers. Statistical inference for Hüsler–Reiss graphical models through matrix completions. *Journal of the American Statistical Association*, 120(550):909–921, 2025. doi: 10.1080/01621459.2024.2371978.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. doi: 10.1016/0378-8733(83)90021-7.

Jürg Hüsler and Rolf-Dieter Reiss. Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286, 1989. doi: 10.1016/0167-7152(89)90106-5.

Vassilis Kalofolias. How to learn a graph from smooth signals. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 920–929. PMLR, 2016. URL https://proceedings.mlr.press/v51/kalofolias16.html.

Ioannis Koutis, Alex Levin, and Richard Peng. Improved spectral sparsification and numerical algorithms for SDD matrices. In *Proceedings of the 29th International Symposium on Theoretical Aspects of Computer Science (STACS'12)*, pages 266–277, 2012. doi: 10.4230/LIPIcs.STACS.2012.266.

Sandeep Kumar, Jiaxi Ying, José Vinícius de Miranda Cardoso, and Daniel P. Palomar. A unified framework for structured graph learning via spectral constraints. *Journal of Machine Learning Research*, 21(22):1–60, 2020. URL http://jmlr.org/papers/v21/19-276.html.

Rasmus Kyng and Sushant Sachdeva. Approximate Gaussian elimination for Laplacians – Fast, sparse, and simple. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 573–582. IEEE, 2016. doi: 10.1109/FOCS.2016.68.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 978-0-19-852219-5.

Steffen L. Lauritzen and Piotr Zwiernik. Locally associated graphical models and mixed convex exponential families. *The Annals of Statistics*, 50(5):3009–3038, 2022. doi: 10.1214/22-AOS2219.

Steffen L. Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. *The Annals of Statistics*, 47(4):1835–1863, 2019. doi: 10.1214/17-AOS1668.

Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. *SIAM Journal on Computing*, 47(6):2315–2336, 2018. doi: 10.1137/16M1061850.

Holger Rootzén and Nader Tajvidi. Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930, 2006. doi: 10.3150/bj/1161614952.

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

Veeru Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. Graph sparsification approaches for Laplacian smoothing. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1250–1259. PMLR, 2016. URL https://proceedings.mlr.press/v51/sadhanala16.html.

Martin Slawski and Matthias Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015. doi: 10.1016/j.laa.2014.04.020.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. doi: 10.1137/080734029.

Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC'04)*, pages 81–90, 2004. doi: 10.1145/1007352.1007372.

Nikhil Srivastava and Roman Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *The Annals of Probability*, 41(5):3081–3111, 2013. doi: 10.1214/12-AOP760.

Yongyu Wang, Zhiqiang Zhao, and Zhuo Feng. Scalable graph topology learning via spectral densification. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1099–1108. ACM, 2022. doi: 10.1145/3488560.3498480. URL https://dl.acm.org/doi/10.1145/3488560.3498480.

Yuhao Wang, Uma Roy, and Caroline Uhler. Learning high-dimensional Gaussian graphical models under total positivity without adjustment of tuning parameters. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108 of *Proceedings of Machine Learning Research*, pages 2698–2708. PMLR, 2020. URL https://proceedings.mlr.press/v108/wang20g.html.

Jiaxi Ying, José Vinícius de Miranda Cardoso, and Daniel P. Palomar. Nonconvex sparse graph learning under Laplacian constrained graphical model. In *Advances in Neural Information Processing Systems*, volume 33, pages 7101–7113. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4ef42b32bccc9485b10b8183507e5d82-Paper.pdf.

Jiaxi Ying, José Vinícius de Miranda Cardoso, and Daniel P. Palomar. Minimax estimation of Laplacian constrained precision matrices. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3736–3744. PMLR, 2021. URL https://proceedings.mlr.press/v130/ying21a.html.

Licheng Zhao, Yiwei Wang, Sandeep Kumar, and Daniel P. Palomar. Optimization algorithms for graph Laplacian estimation via ADMM and MM. *IEEE Transactions on Signal Processing*, 67(16):4231–4244, 2019. doi: 10.1109/TSP.2019.2925602.

## Appendix A. Proofs

### A.1. Bregman–Loewner Lemmas

The proof of Theorem 3.1 (which, applied under the substitution of Section 4.1, also covers the Hüsler–Reiss case) and of Corollary A.4 reduce to two elementary lemmas about graph Laplacians sharing the kernel $\operatorname{span}(\mathbf{1})$. The full-rank versions of these lemmas, for positive definite matrices, were proved in Echave-Sustaeta Rodríguez et al. (2026, Sec. 3). The two statements below adapt them to the degenerate-rank setting that arises here for both LCGGMs and the Hüsler–Reiss surrogate.

**Lemma A.1** *Let $K, \widetilde{K}$ be positive semidefinite matrices with common kernel $\operatorname{span}(\mathbf{1})$, and suppose $(1-\varepsilon)K \preceq \widetilde{K} \preceq (1+\varepsilon)K$ on $\mathbf{1}^\perp$ for some $\varepsilon \in (0,1)$. Then $D_{\mathrm{KL}}^{\perp}(K \,\|\, \widetilde{K}) \geq 0$, and*

$$2D_{\mathrm{KL}}^{\perp}(K \,\|\, \widetilde{K}) \;\leq\; \frac{(d-1)\varepsilon^2}{2(1-\varepsilon)}. \tag{17}$$

*In particular, $2D_{\mathrm{KL}}^{\perp} \leq (d-1)\varepsilon^2$ for $\varepsilon \leq \frac{1}{2}$.*

**Proof** Working on $\mathbf{1}^\perp$, let $\mu_1, \dots, \mu_{d-1}$ be the eigenvalues of $\widetilde{K}K^+$. The spectral approximation gives $\mu_i \in [1-\varepsilon, 1+\varepsilon]$. The standard KL identity for degenerate centered Gaussians on $\mathbf{1}^\perp$ is $2D_{\mathrm{KL}}^\perp(K \,\|\, \widetilde{K}) = \sum_{i=1}^{d-1}(\mu_i - 1 - \log \mu_i) \geq 0$. Writing $\mu_i = 1 + t_i$ with $|t_i| \leq \varepsilon < 1$ and applying the bound $t - \log(1+t) \leq t^2/(2(1-|t|))$ for $|t| < 1$ (Taylor with integral remainder) yields $\mu_i - 1 - \log \mu_i \leq \varepsilon^2/(2(1-\varepsilon))$. Summing gives (17). The simpler bound $(d-1)\varepsilon^2$ follows for $\varepsilon \leq \frac{1}{2}$ since $1/(2(1-\varepsilon)) \leq 1$. ■

**Lemma A.2** *Let $K, \widetilde{K}$ be positive semidefinite matrices with common kernel* $\mathrm{span}(\mathbf{1})$ *and $-\varepsilon K \preceq \widetilde{K} - K \preceq \varepsilon K$ on $\mathbf{1}^\perp$. Then for every symmetric matrix $A$ supported on $\mathbf{1}^\perp$,*

$$\big|\mathrm{tr}\big((\widetilde{K} - K)A\big)\big| \;\leq\; \varepsilon \,\big\|K^{1/2} A K^{1/2}\big\|_*,$$

*where $K^{1/2}$ is the symmetric square root of $K$ on $\mathbf{1}^\perp$.*

**Proof** Let $V \in \mathbb{R}^{d \times (d-1)}$ have orthonormal columns spanning $\mathbf{1}^\perp$ and write $\bar{X} := V^\top X V$ for any symmetric matrix $X$ supported on $\mathbf{1}^\perp$. Then $\mathrm{tr}(XY) = \mathrm{tr}(\bar{X}\bar{Y})$ for such matrices, and $\bar{K} \succ 0$. The spectral approximation is equivalent to $\|\bar{K}^{-1/2}(\bar{\widetilde{K}} - \bar{K})\bar{K}^{-1/2}\|_{\mathrm{op}} \leq \varepsilon$. The operator–trace Hölder inequality gives

$$\big|\mathrm{tr}\big((\widetilde{K} - K)A\big)\big| = \big|\mathrm{tr}\big(\bar{K}^{-1/2}(\bar{\widetilde{K}} - \bar{K})\bar{K}^{-1/2} \cdot \bar{K}^{1/2}\bar{A}\bar{K}^{1/2}\big)\big| \leq \varepsilon \,\big\|\bar{K}^{1/2}\bar{A}\bar{K}^{1/2}\big\|_*,$$

and $\|\bar{K}^{1/2}\bar{A}\bar{K}^{1/2}\|_* = \|K^{1/2}AK^{1/2}\|_*$. ■

### A.2. Proofs for Section 3

**Lemma A.3 (Degenerate-Gaussian KL)** *Let $K, \widetilde{K}$ be positive semidefinite matrices with common kernel* $\mathrm{span}(\mathbf{1})$*. The KL divergence between the two centered degenerate Gaussians on $\mathbf{1}^\perp$ with precisions $K$ and $\widetilde{K}$ is*

$$D_{\mathrm{KL}}^\perp(K \,\|\, \widetilde{K}) \;=\; \tfrac{1}{2}\big[\mathrm{tr}(\widetilde{K}K^+) - (d-1) - \log \mathrm{Det}(\widetilde{K}K^+)\big] \;=\; \tfrac{1}{2}\sum_{i=1}^{d-1}(\mu_i - 1 - \log \mu_i), \quad (18)$$

*where $\mu_1, \dots, \mu_{d-1}$ are the eigenvalues of $\widetilde{K}K^+$ on $\mathbf{1}^\perp$ (equivalently, the generalized eigenvalues of $\widetilde{K}$ relative to $K$ on $\mathbf{1}^\perp$). In particular, $D_{\mathrm{KL}}^\perp(K \,\|\, \widetilde{K}) \geq 0$ with equality iff $\widetilde{K} = K$ on $\mathbf{1}^\perp$.*

**Proof** On $\mathbf{1}^\perp$, both Gaussians are absolutely continuous with respect to the $(d-1)$-dimensional Lebesgue measure on that hyperplane, with densities

$$f_K(y) = \frac{(\mathrm{Det}K)^{1/2}}{(2\pi)^{(d-1)/2}} \exp\big(-\tfrac{1}{2}y^\top K y\big), \quad f_{\widetilde{K}}(y) = \frac{(\mathrm{Det}\widetilde{K})^{1/2}}{(2\pi)^{(d-1)/2}} \exp\big(-\tfrac{1}{2}y^\top \widetilde{K} y\big).$$

Both densities live on the same measure space because $K$ and $\widetilde{K}$ share the kernel $\mathrm{span}(\mathbf{1})$. The log-density difference equals $\frac{1}{2}\log \mathrm{Det}K - \frac{1}{2}\log \mathrm{Det}\widetilde{K} - \frac{1}{2}y^\top(K - \widetilde{K})y$. Taking expectation under $Y \sim f_K$ with $\mathbb{E}_K[YY^\top] = K^+$,

$$\mathbb{E}_K[Y^\top(K - \widetilde{K})Y] = \mathrm{tr}((K - \widetilde{K})K^+) = \mathrm{tr}(KK^+) - \mathrm{tr}(\widetilde{K}K^+) = (d-1) - \mathrm{tr}(\widetilde{K}K^+),$$

using $KK^+ = P$, the projector onto $\mathbf{1}^\perp$ with $\mathrm{tr}(P) = d-1$. Substituting,

$$D_{\mathrm{KL}}^{\perp}(K \,\|\, \widetilde{K}) = \tfrac{1}{2}\big[\mathrm{tr}(\widetilde{K}K^+) - (d-1) + \log \mathrm{Det} K - \log \mathrm{Det}\widetilde{K}\big].$$

Since $K$ and $\widetilde{K}$ share kernel span$(\mathbf{1})$, both are (pseudo-) invertible on $\mathbf{1}^\perp$ and the determinant product rule on $\mathbf{1}^\perp$ gives $\log \mathrm{Det} K - \log \mathrm{Det}\widetilde{K} = -\log \mathrm{Det}(\widetilde{K}K^+)$, yielding the first equality in (18). For the second, let $\mu_i$ be the eigenvalues of $\widetilde{K}K^+$ on $\mathbf{1}^\perp$, then $\mathrm{tr}(\widetilde{K}K^+) = \sum_i \mu_i$, $\log \mathrm{Det}(\widetilde{K}K^+) = \sum_i \log \mu_i$, and $d-1 = \sum_i 1$, so the three terms combine into $\sum_i(\mu_i - 1 - \log \mu_i)$. Non-negativity, with equality iff every $\mu_i = 1$, follows from $\mu - 1 - \log \mu \geq 0$ for $\mu > 0$. ■

**Proof** [Proof of Theorem 3.1] By the KL identity for degenerate Gaussians on $\mathbf{1}^\perp$,

$$2D_{\mathrm{KL}}^{\perp}(\widehat{K} \,\|\, \widetilde{K}) = \mathrm{tr}(\widetilde{K}\widehat{K}^+) - (d-1) + \log \mathrm{Det}\widehat{K} - \log \mathrm{Det}\widetilde{K},$$

so $\log \mathrm{Det}\widetilde{K} - \log \mathrm{Det}\widehat{K} = \mathrm{tr}(\widetilde{K}\widehat{K}^+) - (d-1) - 2D_{\mathrm{KL}}^{\perp}(\widehat{K} \,\|\, \widetilde{K})$. Since $\widehat{K}\widehat{K}^+ = P$ is the projector onto $\mathbf{1}^\perp$, $\mathrm{tr}(\widehat{K}\widehat{K}^+) = d-1$ and $\mathrm{tr}(\widetilde{K}\widehat{K}^+) - (d-1) = \mathrm{tr}((\widetilde{K} - \widehat{K})\widehat{K}^+)$. Substituting into $\ell(\widetilde{K}; T) - \ell(\widehat{K}; T) = (\log \mathrm{Det}\widetilde{K} - \log \mathrm{Det}\widehat{K}) - \mathrm{tr}((\widetilde{K} - \widehat{K})T)$ gives

$$\ell(\widetilde{K}; T) - \ell(\widehat{K}; T) = -2D_{\mathrm{KL}}^{\perp}(\widehat{K} \,\|\, \widetilde{K}) + \mathrm{tr}\big((\widetilde{K} - \widehat{K})(\widehat{K}^+ - T)\big),$$

which is (9). The upper bound in (10) follows by dropping the non-positive KL term and applying Lemma A.2 with $A = T - \widehat{K}^+$. For the lower bound, combine Lemma A.1 ($0 \leq 2D_{\mathrm{KL}}^{\perp} \leq (d-1)\varepsilon^2/(2(1-\varepsilon))$ for $\varepsilon \in (0,1)$) with Lemma A.2. ■

**Corollary A.4** *Run Algorithm 1 on $S$, so that Step 1 returns the LCGGM-MLE $\widehat{K}$ of (4), Step 2 returns a sparsifier $\widetilde{K}$ with edge set $\widetilde{E}$ satisfying (8) for some $\varepsilon \in (0,1)$, and Step 3 returns $\widetilde{K}^{\mathrm{refit}}$. Then*

$$\ell(\widetilde{K}^{\mathrm{refit}}; S) \;\geq\; \ell(\widetilde{K}; S) \;\geq\; \ell(\widehat{K}; S) - \frac{(d-1)\varepsilon^2}{2(1-\varepsilon)} - \varepsilon\, R_{\widehat{K}}(S).$$

*The same holds for Spectral-HR under the substitution of Section 4.1, with $S$ replaced by $\overline{\Sigma}$.*

**Proof** [Proof of Corollary A.4] By (8), $z^\top \widetilde{K} z \geq (1-\varepsilon) z^\top \widehat{K} z > 0$ for every $z \in \mathbf{1}^\perp \setminus \{0\}$, so $\widetilde{K}$ is positive definite on $\mathbf{1}^\perp$ and hence lies in $\mathcal{C}$, and its support is $\widetilde{E}$ by construction. Thus $\widetilde{K}$ is feasible for the Step-3 program, which yields the first inequality. The second is Theorem 3.1 applied at $T = S$. ■

### A.3. Proofs for Section 4

The Hüsler–Reiss Bregman identity and log-likelihood bound quoted in Section 4.1 are (9) and (10), read under the substitution $\widehat{K} \to \widehat{\Theta}$, $\widetilde{K} \to \widetilde{\Theta}$, $S \to \overline{\Sigma}$, $\widehat{K}^+ \to \widehat{\Sigma}$. They are therefore immediate from Theorem 3.1.

**Remark A.5 (Sharpness of the hypothesis of Corollary 4.2)** The condition $\tau_\varepsilon M_\Theta < 1$ is never automatically satisfied. Indeed, $\Theta$ is a principal block of $\mathrm{CM}(\Gamma)^{-1}$, so $M_\Theta \geq \|\Theta\|_{\mathrm{op}} = 1/\mu$ with $\mu := \min\{z^\top \Theta^+ z : z \in \mathbf{1}^\perp, \|z\|_2 = 1\}$, while $\Gamma_{ij} = (e_i - e_j)^\top \Theta^+ (e_i - e_j) \geq 2\mu$ for every $i \neq j$ gives $\|\Gamma\|_{\mathrm{op}} \geq \max_{i \neq j} \Gamma_{ij} \geq 2\mu$. Hence $\|\Gamma\|_{\mathrm{op}} M_\Theta \geq 2$ and $\tau_\varepsilon M_\Theta \geq \varepsilon/(1-\varepsilon)$, so Corollary 4.2 is in force only for $\varepsilon < 1/2$, that is, by Theorem 2.1, only for $\eta > 7 + 4\sqrt{3} \approx 13.9$. Among the fits reported in Section 4.2 this covers the AIC-selected $\eta = 25$ on the Danube data and both selections on the Texas Cluster ($\eta = 20$ and $\eta = 100$), but not the BIC-selected $\eta = 4$ on the Danube data, for which $\varepsilon = 0.8$.

**Proof** [Proof of Theorem 4.1] On $\mathbf{1}^\perp$ both Laplacians are positive definite. The spectral approximation together with the fact that inversion is order-reversing on positive definite operators gives $\Theta^+/(1+\varepsilon) \preceq \widetilde{\Theta}^+ \preceq \Theta^+/(1-\varepsilon)$ on $\mathbf{1}^\perp$. For each pair $i \neq j$, the vector $b_{ij} := e_i - e_j$ lies in $\mathbf{1}^\perp$, so $\widetilde{\Gamma}_{ij} = b_{ij}^\top \widetilde{\Theta}^+ b_{ij}$ inherits the two-sided bound, giving (12). The same argument applies to any pair, so all variogram entries are preserved multiplicatively. ■

**Proof** [Proof of Corollary 4.2] Write $C := \mathrm{CM}(\Gamma)$, $\widetilde{C} := \mathrm{CM}(\widetilde{\Gamma})$ and $v(y) := \binom{y}{1} \in \mathbb{R}^{d+1}$. By the Hüsler–Reiss density representation (6),

$$\log \lambda_{\widetilde{\Gamma}}(y) - \log \lambda_\Gamma(y) = -\tfrac{1}{2}\big(\log|\det \widetilde{C}| - \log|\det C|\big) - \tfrac{1}{2}\, v(y)^\top (\widetilde{C}^{-1} - C^{-1})\, v(y).$$

*Step 1 (Cayley–Menger perturbation).* The Cayley–Menger blocks yield

$$\widetilde{C} - C = \begin{pmatrix} -(\widetilde{\Gamma} - \Gamma)/2 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \text{so} \qquad \|\widetilde{C} - C\|_{\mathrm{op}} = \tfrac{1}{2}\|\widetilde{\Gamma} - \Gamma\|_{\mathrm{op}}.$$

The two-sided bound (12) of Theorem 4.1 reads $\Gamma_{ij}/(1+\varepsilon) \leq \widetilde{\Gamma}_{ij} \leq \Gamma_{ij}/(1-\varepsilon)$, which translates to the entrywise bound $|\widetilde{\Gamma}_{ij} - \Gamma_{ij}| \leq \frac{\varepsilon}{1-\varepsilon}\Gamma_{ij}$ (the upper deviation $\varepsilon/(1-\varepsilon)$ dominates the lower $\varepsilon/(1+\varepsilon)$). Since $\Gamma$ has nonnegative entries, monotonicity of the spectral radius on nonnegative matrices (Wielandt's theorem) gives $\rho(|\widetilde{\Gamma} - \Gamma|) \leq \frac{\varepsilon}{1-\varepsilon}\rho(\Gamma)$, and the symmetry of both matrices together with $\|A\|_{\mathrm{op}} \leq \rho(|A|)$ for symmetric $A$ gives $\|\widetilde{\Gamma} - \Gamma\|_{\mathrm{op}} \leq \rho(|\widetilde{\Gamma} - \Gamma|) \leq \frac{\varepsilon}{1-\varepsilon}\|\Gamma\|_{\mathrm{op}}$. Hence $\|\widetilde{C} - C\|_{\mathrm{op}} \leq \tau_\varepsilon$. Using $\|\Gamma\|_{\mathrm{op}}$ here, instead of the entrywise Frobenius bound $\|\widetilde{\Gamma} - \Gamma\|_F \leq \frac{\varepsilon}{1-\varepsilon}\|\Gamma\|_F$ that an entry-by-entry summation would give, gains a constant factor: for $\Gamma \in \mathcal{D}^d$ the ratio $\|\Gamma\|_F/\|\Gamma\|_{\mathrm{op}}$ is at most $\sqrt{2}$, since such a $\Gamma$ has zero trace and exactly one positive eigenvalue, which therefore equals the sum of the absolute values of the remaining ones.

*Step 2 (log-determinant and inverse perturbation).* Write $M_\Theta := \|C^{-1}\|_{\mathrm{op}} = 1/s_{\min}(C)$, so $s_i(C) \geq s_{\min}(C) = 1/M_\Theta$ for every $i$. Assuming $\tau_\varepsilon M_\Theta < 1$, Weyl's inequality for singular values ($|s_i(\widetilde{C}) - s_i(C)| \leq \|\widetilde{C} - C\|_{\mathrm{op}} \leq \tau_\varepsilon$) gives

$$s_i(\widetilde{C}) \geq s_i(C) - \tau_\varepsilon \geq 1/M_\Theta - \tau_\varepsilon = (1 - \tau_\varepsilon M_\Theta)/M_\Theta > 0 \quad \text{for every } i,$$

and, writing $t_i := s_i(\widetilde{C})/s_i(C) - 1$, $|t_i| \leq \tau_\varepsilon/s_i(C) \leq \tau_\varepsilon M_\Theta < 1$. Using $|\log(1+t)| \leq |t|/(1-|t|)$ for $|t| < 1$ (immediate from $\log(1+t) = \int_0^1 t/(1+\tau t)\, d\tau$), each term satisfies $|\log(s_i(\widetilde{C})/s_i(C))| = |\log(1+t_i)| \leq |t_i|/(1-|t_i|) \leq \tau_\varepsilon M_\Theta/(1-\tau_\varepsilon M_\Theta)$. Summing over the

$d+1$ singular values of $C$,

$$\big|\log|\det\widetilde{C}| - \log|\det C|\big| = \Big|\sum_{i=1}^{d+1}\log\big(s_i(\widetilde{C})/s_i(C)\big)\Big| \\ \le \sum_{i=1}^{d+1}|\log(1+t_i)| \;\le\; (d+1)\,\tau_\varepsilon M_\Theta/(1-\tau_\varepsilon M_\Theta),$$

where the first equality uses $|\det A| = \prod_i s_i(A)$. For the inverse perturbation, the resolvent identity $\widetilde{C}^{-1} - C^{-1} = C^{-1}(C-\widetilde{C})\widetilde{C}^{-1}$ together with $\|C^{-1}\|_{\mathrm{op}} = M_\Theta$ and $\|\widetilde{C}^{-1}\|_{\mathrm{op}} = 1/s_{\min}(\widetilde{C}) \le M_\Theta/(1-\tau_\varepsilon M_\Theta)$ gives

$$\|\widetilde{C}^{-1} - C^{-1}\|_{\mathrm{op}} \le M_\Theta \cdot \tau_\varepsilon \cdot M_\Theta/(1-\tau_\varepsilon M_\Theta) = \tau_\varepsilon M_\Theta^2/(1-\tau_\varepsilon M_\Theta),$$

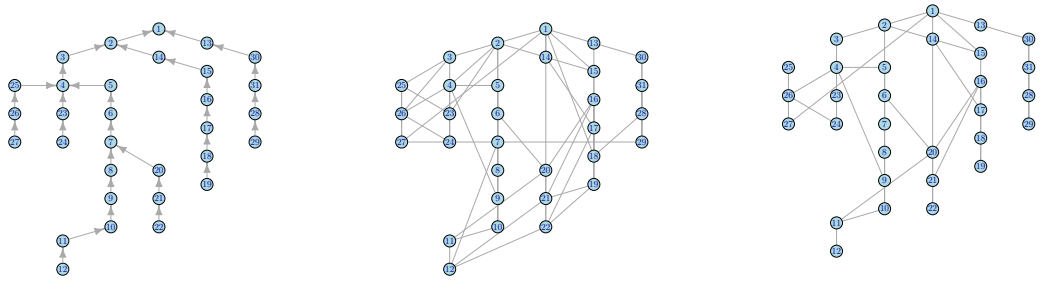
whence $|v^\top(\widetilde{C}^{-1} - C^{-1})v| \le \|v\|_2^2\,\|\widetilde{C}^{-1} - C^{-1}\|_{\mathrm{op}} = (\|y\|_2^2+1)\,\tau_\varepsilon M_\Theta^2/(1-\tau_\varepsilon M_\Theta)$ (using $\|v(y)\|_2^2 = \|y\|_2^2 + 1$). Combining the two bounds yields (15). The block bound $M_\Theta \le \|\Theta\|_{\mathrm{op}} + \|q\|_2 + |\sigma^2|$ in (13) follows by triangle inequality on the three-block decomposition $\mathrm{CM}(\Gamma)^{-1} = \mathrm{diag}(\Theta,0) + \left(\begin{smallmatrix} 0 & q \\ q^\top & 0 \end{smallmatrix}\right) + \mathrm{diag}(0,\sigma^2)$, whose summands have operator norms $\|\Theta\|_{\mathrm{op}}$, $\|q\|_2$, and $|\sigma^2|$. The sample-mean form stated after (15) follows by averaging the pointwise bound over $Y_1,\dots,Y_n$. ∎

## Appendix B. Computational cost of Algorithm 1

Step 1 is a convex program: the objective $\ell(K;S)$ is concave on the cone of connected Laplacians, since $\log\mathrm{Det}$ is the log-determinant of the restriction of $K$ to $\mathbf{1}^\perp$, and the constraints are linear. It can therefore be solved to global optimality by standard first- and second-order methods, for instance block coordinate descent (Egilmez et al., 2017; Röttger et al., 2023), projected gradient descent (Ying et al., 2020), or ADMM (Zhao et al., 2019). For the latter two the per-iteration cost is $O(d^3)$, dominated by a $d \times d$ inversion or eigendecomposition. Our implementation uses an edge-wise coordinate descent in the spirit of Lauritzen and Zwiernik (2022): each edge weight is updated by a projected Newton step, which requires only the current effective resistance of the edge, and the pseudoinverse $K^+$ is maintained by a rank-one update at cost $O(d^2)$ per step, so that a full sweep over all $\binom{d}{2}$ candidate edges costs $O(d^4)$, apart from an $O(d^3)$ refactorization whenever an edge leaves the support (an active-set strategy sweeps only the current support in later iterations). Step 2, the BSS construction, performs $\lceil \eta(d-1)\rceil$ rounds, each of which solves two $(d-1)\times(d-1)$ linear systems and scores all $m$ edges of $\widehat{K}$, for a total of $O(\eta\, d^3 m)$ operations, with $m \le \binom{d}{2}$. Almost-linear-time constructions of linear-size sparsifiers are also available (Lee and Sun, 2018). Step 3 is Step 1 restricted to at most $\lceil \eta(d-1)\rceil$ edges, which is considerably cheaper than the dense fit.

## Appendix C. Estimated graphs for the Danube experiment

Figure 1 displays the two fits discussed in Section 4.2 next to the physical river network, on a common layout.



(a) Flow graph

(b) $\mathrm{EMTP}_2$ (67 edges)

(c) Spectral-HR, $\eta = 4$ (46 edges)

Figure 1: Danube river network: (a) the directed flow graph from Engelke and Hitz (2020), (b) the $\mathrm{EMTP}_2$-MLE fit, and (c) the Spectral-HR fit at the BIC-selected $\eta = 4$. Node positions and numbering match across panels.

## Appendix D. Tuning-parameter sweeps for the Danube experiment

(a) Spectral-HR ($\eta$)

| $\eta$ | Edges | BIC | AIC |
|---|---|---|---|
| 1.10 | 31 | 1,521.13 | 1,435.50 |
| 1.20 | 33 | 1,364.67 | 1,273.51 |
| 1.30 | 34 | 1,342.50 | 1,248.59 |
| 1.50 | 35 | 1,304.65 | 1,207.97 |
| 2.00 | 38 | 1,285.99 | 1,181.03 |
| 2.50 | 39 | 1,281.33 | 1,173.60 |
| 3.00 | 42 | 1,284.26 | 1,168.25 |
| 4.00 | 46 | **1,264.51** | 1,137.45 |
| 5.00 | 46 | 1,264.51 | 1,137.45 |
| 6.00 | 48 | 1,268.56 | 1,135.97 |
| 8.00 | 48 | 1,268.56 | 1,135.97 |
| 10.00 | 48 | 1,268.56 | 1,135.97 |
| 15.00 | 50 | 1,265.84 | 1,127.73 |
| 20.00 | 52 | 1,268.50 | 1,124.87 |
| 25.00 | 53 | 1,268.74 | **1,122.34** |
| 30.00 | 53 | 1,268.74 | 1,122.34 |
| 40.00 | 55 | 1,277.54 | 1,125.62 |
| 50.00 | 56 | 1,281.83 | 1,127.15 |
| 75.00 | 57 | 1,287.78 | 1,130.33 |
| 100.00 | 60 | 1,302.27 | 1,136.54 |

(b) eglearn-NS ($\rho$)

| $\rho$ | Edges | BIC | AIC |
|---|---|---|---|
| 0.005 | 364 | 2,012.64 | 1,007.21 |
| 0.014 | 184 | 1,342.89 | 834.65 |
| 0.023 | 130 | 1,179.38 | **820.29** |
| 0.032 | 99 | **1,176.95** | 903.49 |
| 0.041 | 89 | 1,281.86 | 1,036.03 |
| 0.050 | 82 | 1,309.73 | 1,083.23 |
| 0.091 | 56 | 1,519.44 | 1,364.76 |
| $\geq 0.132$ | — | — | — |

(c) eglearn-glasso ($\rho$)

| $\rho$ | Edges | BIC | AIC |
|---|---|---|---|
| 0.005 | 291 | 1,694.25 | 890.46 |
| 0.044 | 171 | **1,342.75** | **870.42** |
| 0.083 | 151 | 1,419.67 | 1,002.58 |
| 0.123 | 146 | 1,531.29 | 1,128.01 |
| 0.162 | 160 | 1,538.55 | 1,096.60 |
| 0.201 | 167 | 1,556.76 | 1,095.48 |
| 0.240 | 174 | 1,577.93 | 1,097.31 |
| 0.279 | 171 | 1,571.70 | 1,099.36 |
| 0.319 | 170 | 1,543.68 | 1,074.11 |
| 0.358 | 170 | 1,557.99 | 1,088.42 |
| 0.397 | 162 | 1,566.81 | 1,119.34 |
| 0.436 | 158 | 1,544.02 | 1,107.60 |
| 0.476 | 150 | 1,567.51 | 1,153.18 |
| 0.515 | 134 | 1,561.81 | 1,191.68 |
| 0.554 | 128 | 1,709.40 | 1,355.84 |
| 0.593 | 117 | 1,707.51 | 1,384.33 |
| $\geq 0.632$ | — | — | — |

Table 3: Full tuning-parameter sweeps on the Danube data for the three tunable methods: edge count, BIC and AIC (16) at every grid point. (a) Spectral-HR over the BSS parameter $\eta$, (b) eglearn-NS over the $\ell_1$ parameter $\rho$, (c) eglearn-glasso over $\rho$. Bold marks each criterion's optimum over the grid, that is the row selected by that criterion, where consecutive grid points yield the same support, and hence the same refitted model, the smallest one is marked. Rows listed as "—" (in (b) for $\rho \geq 0.132$ and in (c) for $\rho \geq 0.632$) correspond to grid points where the estimated graph is disconnected and $\widehat{\Gamma}$ cannot be completed.

## Appendix E. Texas Cluster of U.S. airports

This appendix gives the full version of the second real-data experiment of Section 4.2.

The data are daily total flight delays at $d = 29$ airports, accessed via the `graphicalExtremes` R package (Engelke et al., 2024), with the predefined train/test date split of Hentschel et al. (2025) ($n_{\text{train}} = 1{,}764$, $n_{\text{test}} = 1{,}839$ days). At threshold $p = 0.85$ this yields $n_{\text{tr}}^{\text{thr}} = 1{,}069$ training and $n_{\text{te}}^{\text{thr}} = 1{,}336$ held-out exceedances. Again there is no canonical ground-truth graph, so we evaluate each fit by edge count, training Hüsler–Reiss BIC and AIC (16), and held-out Hüsler–Reiss log-likelihood, and each tunable method is selected twice. The eglearn grid is the 17-point union of 6 equidistant values on $[0.005, 0.05]$ and 12 on $[0.05, 0.5]$, finer where the fits change fastest, and the Spectral-HR grid is 24 values of $\eta$ spanning $[1.1, 500]$. Both selected parameters are interior to their grids, and the $\eta$ grid runs well past the range needed for BIC because the weaker AIC penalty selects a considerably denser fit.

Table 4 reports both selections. Under BIC, Spectral-HR gives the sparsest fit by a clear margin (95 edges, vs 130 for $\text{EMTP}_2$ and 211–227 for the eglearn variants). $\text{EMTP}_2$ attains the best held-out test log-likelihood, with Spectral-HR a close second ($-23{,}104$ vs

−23,209) at a 27% reduction in edge count. The eglearn variants achieve the best training BIC but visibly overfit: their held-out log-likelihoods (−23,900 for eglearn-NS and −23,851 for eglearn-glasso) are substantially worse than both $\mathrm{EMTP}_2$ and Spectral-HR despite the lower training criterion.

Because the AIC penalty $2k$ is far weaker than BIC's $k \log n_{\mathrm{tr}}^{\mathrm{thr}}$, every method is selected denser under AIC, and the two criteria separate the methods further. The denser eglearn fits are worse out of sample (−24,048 at 242 edges for eglearn-NS, −24,028 at 264 for eglearn-glasso): both already overfit at their BIC optimum, and AIC moves them further in. The denser Spectral-HR fit instead *improves* the held-out log-likelihood, to −23,138 at 110 edges, within 34 nats of the dense $\mathrm{EMTP}_2$ baseline at 85% of its edges. On this dataset the sparsification path thus degrades gracefully where the $\ell_1$ paths do not, and it is insensitive to the exact $\eta$: over $\eta \in [50, 200]$ the fit stays within 3 AIC points of its optimum at 106–117 edges.

| Method | Selected by | Param | Edges | Train BIC | Train AIC | HR test log-lik. | Time (s) |
|---|---|---|---|---|---|---|---|
| $\mathrm{EMTP}_2$ | (none) | – | 130 | 38,136.2 | 37,489.5 | **−23,104.3** | 0.4 |
| eglearn-NS | BIC | 0.032 | 211 | **37,289.2** | 36,239.6 | −23,899.8 | 62.0 |
| | AIC | 0.023 | 242 | 37,349.4 | 36,145.6 | −24,047.6 | |
| eglearn-glasso | BIC | 0.050 | 227 | 37,369.3 | 36,240.1 | −23,851.5 | 59.4 |
| | AIC | 0.032 | 264 | 37,454.0 | **36,140.7** | −24,027.7 | |
| Spectral-HR | BIC | 20 | **95** | 37,967.3 | 37,494.7 | −23,208.5 | 99.2 |
| | AIC | 100 | 110 | 38,013.9 | 37,466.7 | −23,138.2 | |

Table 4: Texas Cluster of U.S. airports, daily total flight delays (Hüsler–Reiss, $d = 29$), with the train/test split and thresholds given in the text. Tuning on the training sample by Hüsler–Reiss BIC and AIC (16), both evaluated at $n_{\mathrm{tr}}^{\mathrm{thr}} = 1{,}069$ and computed from a single log-likelihood evaluation, so that their difference is exactly $k(\widehat{\Theta})\,(\log n_{\mathrm{tr}}^{\mathrm{thr}} - 2)$. As in Table 2, each tunable method appears once per criterion. HR test log-lik. is the held-out Hüsler–Reiss log-likelihood, the headline metric here. Time is the wall-clock time in seconds for the complete sweep of each method over its tuning grid (17 values of $\rho$ for eglearn, 24 values of $\eta$ for Spectral-HR), shared by both selections. Bold: best on each metric across all rows, timings excluded.