




Learning Networks from Gaussian Graphical Models and Gaussian Free Fields

Subhro Ghosh¹ · Soumendu Sundar Mukherjee² · Hoang-Son Tran¹  · Ujan Gangopadhyay¹

Received: 26 September 2023 / Accepted: 4 March 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

We investigate the problem of estimating the structure of a weighted network from repeated measurements of a Gaussian graphical model (GGM) on the network. In this vein, we consider GGMs whose covariance structures align with the geometry of the weighted network on which they are based. Such GGMs have been of longstanding interest in statistical physics, and are referred to as the Gaussian free field (GFF). In recent years, they have attracted considerable interest in the machine learning and theoretical computer science. In this work, we propose a novel estimator for the weighted network (equivalently, its Laplacian) from repeated measurements of a GFF on the network, based on the Fourier analytic properties of the Gaussian distribution. In this pursuit, our approach exploits complex-valued statistics constructed from observed data, that are of interest in their own right. We demonstrate the effectiveness of our estimator with concrete recovery guarantees and bounds on the required sample complexity. In particular, we show that the proposed statistic achieves the parametric rate of estimation for fixed network size. In the setting of networks growing with sample size, our results show that for Erdos–Renyi random graphs $G(d, p)$ above the connectivity threshold, network recovery takes place with high probability as soon as the sample size n satisfies $n \gg d^4 \log d \cdot p^{-2}$.

Keywords Precision matrix · Gaussian free field · Gaussian graphical model

Mathematics Subject Classification Primary 62F12 · Secondary 62F10 · 62F35

Communicated by Federico Ricci-Tersenghi.

Subhro Ghosh, Soumendu Sundar Mukherjee and Hoang-Son Tran have contributed equally to this work.

✉ Hoang-Son Tran
hoangson.tran@u.nus.edu

1 Introduction

1.1 Gaussian Graphical Models

Gaussian graphical models (GGM), also known as Gaussian Markov random fields (GMRF), are multivariate Gaussian distributions defined on undirected graphs. In these models, a Gaussian random variable is associated to each vertex of a graph, and the existence or non-existence of edges captures the dependency structure between these random variables.

More precisely, suppose we have a graph $G = (V, E)$ on $|V| = d$ vertices. In a GMRF/GGM model, we assume our sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ consists of i.i.d. copies of a random vector $\mathbf{X} = (X_1, \dots, X_d)^\top$ having a multivariate normal distribution.

It is well-known that the covariance matrix Σ of a GGM captures the dependency structure between X_i s. That is, for $i \neq j$, the (i, j) -th matrix entry $\Sigma[i, j] = 0$ if and only if the variable X_i is independent of the variable X_j .

The inverse of the covariance matrix, often called the *precision matrix* or the information matrix, is also of longstanding interest, because it captures the conditional dependency structure between X_i s. More precisely, the (i, j) -th matrix entry $\Sigma^{-1}[i, j]$ for $i \neq j$ is 0 if and only if X_i is independent of X_j given the rest of the variables.

More generally, the following may be shown to hold true. If, for some subset of indices $A \subset V$, it holds that removing the vertices in A from the graph G disconnects the vertices i and j , then the random variables X_i and X_j are conditionally independent given $(X_k)_{k \in A}$. This is a form of graphical *Markov property*, which endows GGM-s with a highly attractive structure, both as a stochastic model and as a data modelling tool.

Unsurprisingly, GGM-s have attracted interest in a wide array of application domains as an effective modelling technique to capture the dependency structures among variates. These include applications to genomics (Menéndez et al. [37], Basso et al. [6], Wille et al. [52], Schafer and Strimmer [44]); neuroscience (Huang et al. [27], Varoquaux et al. [49], Rish et al. [42], Varoquaux et al. [48]); causal inference (Loh and Bühlmann [33]); to name a few.

1.2 Learning Networks from Random Fields

The problem of learning a network from observations of a random field that lives on it has been a topic of great interest in recent years. A significant instance of this is accorded by the Ising model on graphs and Gaussian Graphical Models. The Ising model on a graph G is a random field with values in the set $\{+1, -1\}$, with a dependency structure that reflects the structure of the graph. Estimating the underlying graph (or various properties thereof) from observations of the Ising model (or GGM) has been a topic of intensive research activities in recent years. Investigations on this problem have been carried out in various settings, for details we refer the interested reader to (Ravikumar et al. [40], Bresler [13], Bhattacharya and Mukherjee [10], Berthet et al. [9], Anandkumar et al. [1, 2]) for a partial list of references.

In this paper, we investigate the problem of learning a network from a Gaussian Graphical Model supported on the network. More generally than unweighted graphs, we will concern ourselves with the broader problem of estimating a weighted network from a GGM supported thereon. This necessitates the correlation structure of the GGM to carry information about the weights on the edges of the network. A canonical choice for such a GGM is proffered by the so-called *Gaussian Free Field* (abbrv. GFF), which is what we will focus on in this work.

1.3 Gaussian Free Fields

Gaussian Free Fields (abbrv. GFF) have emerged as important models of strongly correlated Gaussian fields, that are canonically equipped to capture the geometry of their ambient space. In the case of graphs, the background geometry is encapsulated in the graph Laplacian. GFF-s arise originally in theoretical physics, in the study Euclidean quantum field theories. Applications in physics generally require the GFF to be defined on the continuum, which is often a challenge in context of the fact that, even on Euclidean spaces of dimension > 1 , the continuum GFF is defined only as a distribution valued random variable. On graphs and weighted networks, the setting we are principally interested in, it is of interest to study the so-called *Discrete Gaussian Free Field* (abbrv. DGFF). The DGFF, in comparison to the continuum setting, is well-defined as a random field that lives on the nodes of the network; however, it exhibits degeneracy properties as a Gaussian random vector, which demands some consideration in its definition.

The GFF is, for many natural reasons, a GGM of wide interest in its own right. For one, a quadratic form based on the network Laplacian encodes smoothness with respect to the geometry of the weighted network. This underpins the significance of GFF based models in active and semi-supervised learning (Zhu et al. [58, 59], Ma et al. [34], Ghosh and Mukherjee [25]). In fact, it has been shown (Kelner et al. [29]) that DGFFs essentially cover all possibilities in an important class of GGMs known as *attractive* GGMs, wherein the pairwise correlations are all non-negative and which arise naturally in a wide array of applications such as phylogenetic studies and copula models of finance. For more details on the generalities of the GFF and its significance with statistical and mathematical physics, we refer the reader to the excellent mathematical surveys (Sheffield [45], Berestycki [8]).

1.3.1 The Massless DGFF

The massless DGFF on a weighted graph $G = (V(G), E(G))$ is defined as follows. Let ∂ be a distinguished set of vertices, called the boundary of the graph. Let S_n be the simple symmetric random walk on G . Let τ be the hitting time of ∂ . The Green function $\mathbf{G}(x, y)$ is defined for $x, y \in V(G)$ by putting

$$\mathbf{G}(x, y) = \frac{1}{\text{deg}(y)} \mathbb{E}_x \left(\sum_{n=0}^{\infty} \mathbb{1}[X_n = y; \tau > n] \right).$$

The DGFF is the centered Gaussian vector $(h(x))_{x \in V(G)}$ with covariance given by the Green function \mathbf{G} . In other words, if $A \subset \mathbb{R}^{|V(G)|}$, then the probability distribution of $(h(x))_{x \in V(G)}$ is given by (see Theorem 1.5 in Berestycki [8])

$$\mathbb{P} \left((h(x))_{x \in V(G)} \in A \right) = \frac{1}{Z} \int_A \exp \left(-\frac{1}{4} \sum_{x_i \sim x_j} (h(x_i) - h(x_j))^2 \right) \prod_{x_i \notin \partial} dh(x_i) \prod_{x_i \in \partial} \delta_0(dh(x_i)), \tag{1}$$

where δ_0 is the Dirac delta measure at 0. In particular, the field always takes value 0 at the boundary vertices ∂ .

1.3.2 The Massive DGFF

The massive DGFF on a weighted graph $G = (V(G), E(G))$ is, in a sense, simpler than the massless case, and is defined as follows. For the *mass parameter* μ , the massive DGFF

$(h(x))_{x \in V(G)}$ on G is given by the following relation. For $A \subset \mathbb{R}^{|V(G)|}$, then the probability distribution of $(h(x))_{x \in V(G)}$ is given by

$$\mathbb{P}((h(x))_{x \in V(G)} \in A) = \frac{1}{Z} \int_A \exp\left(-\frac{1}{4} \sum_{x_i \sim x_j} (h(x_i) - h(x_j))^2 - \frac{1}{4} \mu \sum_{x_i \in V} h(x_i)^2\right) \prod_{x_i \in V} dh(x_i). \tag{2}$$

In other words, the probability density function of $\mathbf{h} := (h(x))_{x \in V(G)}$ is given, up to a normalization constant, by

$$\exp\left(-\frac{1}{4} \mathbf{h}^\top (\mathbf{L} + \mu \mathbf{I}) \mathbf{h}\right), \tag{3}$$

where \mathbf{L} is the standard Laplacian matrix of the graph, and μ is the mass parameter of the model.

The massive DGFF is the setting we will concern ourselves with in this article. The mass parameter μ clearly controls the well-conditionedness of the model; and the regime we will particularly concern ourselves with is the one where the parameter μ is not too large, thereby focusing on poorly conditioned GGMs.

1.4 Learning Networks from Gaussian Free Fields

In this article we consider the problem of precision matrix estimation of the Gaussian Free Field. To wit, we consider a graph $G = (V(G), E(G))$ with vertex set $V(G)$ and edge set $E(G)$. Let $|V(G)| = d$ be the number of vertices and we label the vertex set with $\{1, \dots, d\}$. Suppose we have an i.i.d. sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ of random vectors in \mathbb{R}^d where each \mathbf{X}_m follows a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. The covariance matrix is related to the graph G in the following way. We assume $\Sigma = (\mathbf{L} + \mu \mathbf{I})^{-1}$ where \mathbf{L} is the graph Laplacian and $\mu > 0$ is an unknown parameter.

The graph Laplacian \mathbf{L} is the $d \times d$ matrix defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where \mathbf{D} is the degree matrix and \mathbf{A} is the adjacency matrix. Thus, for an unweighted graph, (i, i) -th entry of \mathbf{L} is the degree of vertex i , and for $i \neq j$ the (i, j) -th entry of \mathbf{L} is 0 if i is not connected to j , and it is -1 if i is connected to j . For a weighted graph, \mathbf{A} is the *weighted* adjacency matrix of the graph; i.e. $\mathbf{A}_{ij} = w_{ij}$, where w_{ij} is the weight of the edge between vertices i and j . The matrix \mathbf{D} in this setting would be the weighted degree matrix of the graph; i.e., \mathbf{D} is a diagonal matrix with $\mathbf{D}_{ii} = \sum_{j \in V(G)} \mathbf{A}_{ij}$.

We will concern ourselves with the problem of estimating \mathbf{L} from $\mathbf{X}_1, \dots, \mathbf{X}_n$.

Thus our problem is same as estimating the graph underlying a DGFF where μ can be thought of as the number of vertices that are designated as the boundary. These boundary vertices are connected with all the other vertices.

1.5 A Survey on the Estimation of Precision Matrices

In view of its multifaceted importance, [21] initiated investigations into the problem of estimating the precision matrix of a GGM. In this subsection, we provide a partial survey of the principal approaches to this longstanding problem, and the associated problem of estimating the covariance matrix of a GGM.

1.5.1 Estimation of Covariance Matrices

Estimating the covariance matrix of a GGM is not difficult when the sample size n is much bigger than d . In this case the sample covariance matrix

$$\Sigma_n := \frac{1}{n} \sum_{k=1}^n (\mathbf{X}_k - \bar{\mathbf{X}}_n)(\mathbf{X}_k - \bar{\mathbf{X}}_n)^\top, \text{ where } \bar{\mathbf{X}}_n := \frac{1}{n} \sum_{k=1}^n \mathbf{X}_k,$$

is a natural estimator of Σ , and has good consistency properties. In the high-dimensional setup i.e., when the number of variables d is much larger than n , the sample covariance matrix is not a good estimator of the true covariance matrix Σ . In this case estimating the covariance matrix is a challenging problem, estimating the precision matrix is even more difficult.

1.5.2 Estimation Under Order Structures

It is possible to estimate Σ using Σ_n consistently if there exists additional structure on the variables. For example, if the variables have a certain total ordering, for example in a time series data, then one may assume $\Sigma[i, j]$ is 0 or near 0 when $|i - j|$ is big enough. This leads to a banding structure in Σ . Under this kind of assumptions Bickel and Levina [11] showed that banding the sample covariance matrix leads to a consistent estimator. Cai et al. [14] considered the same class of estimators as Bickel and Levina [11] and established the minimax rate of convergence and also constructed a rate-optimal estimator. The minimax rate is given by

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}_\alpha} \mathbb{E}(\|\hat{\Sigma} - \Sigma\|^2) \asymp \min \left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log d}{n}, \frac{d}{n} \right\}.$$

Here α is a sparsity parameter and \mathcal{P}_α is a class of sparse covariance matrices. Larger α corresponds to sparser matrices.

Assuming certain ordering structures on the variables, methods based on banding the Cholesky factor of the inverse covariance matrix for estimating the covariance matrix have also been proposed and studied (see, e.g., Wu and Pourahmadi [54], Huang et al. [26]).

1.5.3 Estimation Under Structured Sparsity

A natural total ordering on the set of variables is unavailable in many situations. Also finding a suitable basis under which the sample-covariance matrix displays a banding structure is often computationally impractical. To deal with these situations Karoui [22] and Bickel and Levina [12] suggested assuming certain permutation invariant sparsity conditions on the covariance matrix and proposed thresholding the sample covariance matrix for estimation. They obtained rates of convergence for the thresholded sample covariance estimator.

1.5.4 Penalized Likelihood Based Methods

Penalized likelihood based methods are also very popular for estimation of sparse precision matrices. Meinshausen and Bühlmann [36] estimate a sparse precision matrix by fitting a lasso model to each variable, using the others as predictors. $\Sigma^{-1}[i, j]$ is then estimated to be nonzero if either the estimated coefficient of variable i on j or the estimated coefficient of

variable j on i is nonzero. They show that asymptotically, this consistently estimates the set of nonzero elements of Σ^{-1} .

1.5.5 The Graphical Lasso

Several algorithms for the exact maximization of the ℓ_1 -penalized log-likelihood have been proposed in the literature (see for e.g. Yuan and Lin [56], Banerjee and Ghaoui [3], Dahl et al. [19], Friedman et al. [24], Rothman et al. [43], Cai et al. [18], Cai et al. [15], Ravikumar et al. [41].) Friedman et al. [24] introduced the Graphical Lasso estimator which minimizes

$$-\log \det(\Sigma^{-1}) + \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X}_i - \boldsymbol{\mu})$$

subject to the sparsity condition $\sum_{i \neq j} \Sigma^{-1}[i, j] \leq t$, where $t \geq 0$ is a tuning parameter. An equivalent formulation is minimizing

$$-\log \det(\Sigma^{-1}) + \text{trace}(\Sigma^{-1} \Sigma_n) + \lambda \sum_{i \neq j} |\Sigma^{-1}[i, j]|$$

where $\lambda \geq 0$ is a tuning parameter. If Σ^{-1} satisfies various conditions, which typically include an assumption similar to or stronger than the restricted eigenvalue (RE) condition (a condition which, in particular, lower bounds the smallest eigenvalue of any $2d \times 2d$ principal submatrix of Σ where d is the maximum vertex degree) then Graphical Lasso succeeds in recovering the graph structure. Further, under some incoherence assumptions on the precision matrix (stronger than RE), it has been shown by Ravikumar et al. [41] that the sparsity pattern of the precision matrix can be accurately recovered from $O((1/\alpha^2)d^2 \log(n))$ samples; here $\alpha \in (0, 1)$ is the incoherence parameter defined as follows. Suppose $\Theta = \Sigma^{-1}$. Let

$$\Gamma := \nabla_{\Theta'}^2 g(\Theta')|_{\Theta'=\Theta} = \Theta^{-1} \otimes \Theta^{-1}$$

where

$$g(\mathbf{A}) := \begin{cases} -\ln \det \mathbf{A} & \text{if } \mathbf{A} > 0 \\ \infty & \text{otherwise} \end{cases}$$

and \otimes denotes the Kronecker matrix product. Thus Γ is a $p^2 \times p^2$ matrix, indexed by vertex pairs so that $\Gamma[(j, k), (l, m)]$ is the partial derivative

$$\frac{\partial^2 g}{\partial \Theta'_{jk} \partial \Theta'_{lm}}$$

evaluated at Θ . For Gaussian observations this is simply $\text{Cov}(X_j X_k, X_l X_m)$. The mutual incoherence or irrepresentable condition is the following:

$$\|\Gamma_{S^c S} (\Gamma_{SS})^{-1}\|_\infty \leq 1 - \alpha$$

for some $\alpha \in (0, 1)$. Here S is the set of vertex pairs either of the form (i, i) or (i, j) if i is connected to j . This condition imposes control on the influence that the non-edge terms, indexed by S^c , can have on the edge-based terms, indexed by S . A similar condition for the Lasso, with the covariance matrix Σ in the place of Γ , is necessary and sufficient for support recovery using the ordinary Lasso ([36, 47, 51, 57].)

1.5.6 The CLIME Estimator

Another popular estimator is the CLIME (constrained ℓ_1 -minimization for inverse matrix estimation) introduced by Cai et al. [18]. It solves the following optimization problem

$$\text{minimize } \|\Theta\|_1 \text{ such that } \|\widehat{\Sigma}_n \Theta - \mathbf{I}\|_\infty \leq \lambda,$$

where λ is a tuning parameter. The analysis of CLIME uses a condition number assumption. For example, if entries $\Sigma^{-1}[i, j]$ are either zero or bounded away from zero by an absolute constant then CLIME succeeds at structure recovery when given roughly $CM^4 \log d$ samples where $M = \max_{\|\mathbf{u}\|_\infty \leq 1} \|\Sigma^{-1}\mathbf{u}\|_\infty$. Cai et al. [16] obtained minimax rates for precision matrix estimation in the high-dimensional setting. They also proposed a fully data driven estimator called ACLIME based on adaptive constrained ℓ_1 minimization and obtained rate of convergence. A survey of minimax rates for sparse covariance matrix estimation and sparse precision matrix estimation along with rates of convergence of various ℓ_1 -penalized estimators can be found in the expository article Cai et al. [17]. d'Aspremont et al. [20] considers penalizing the number of nonzero terms instead of the ℓ_1 penalty. Liu et al. [32] have showed that for a class of non-Gaussian distribution called nonparanormal distribution, the problem of estimating the graph also can be reduced to estimating the precision matrix. Yuan [55] replaced the lasso selection by a Dantzig-type modification, where first the ratios between the off-diagonal elements $\Sigma^{-1}[i, j]$ and the corresponding diagonal element $\Sigma^{-1}[i, i]$ were estimated for each row i and then the diagonal entries $\Sigma[i, i]$ were obtained given the estimated ratios. Lam and Fan [30], Fan et al. [23] considered penalizing the normal likelihood with a nonconvex penalty in order to reduce the bias of the ℓ_1 penalized estimator.

1.5.7 Information Theoretic Lower Bounds

Misra et al. [38] considers the problem of finding information theoretic lower bound on the sample size for recovering the precision matrix in a sparse GGM. They establish that for a model defined on a sparse graph with p nodes, a maximum degree d and minimum normalized edge strength κ , the necessary number of samples scales at least as $d \log p/\kappa^2$. The parameter κ , called the minimum normalized edge strength, is defined as

$$\kappa := \min_{(i,j) \in E} \frac{\Sigma^{-1}[i, j]}{\sqrt{\Sigma^{-1}[i, i]}\sqrt{\Sigma^{-1}[j, j]}}.$$

They propose an algorithm called degree-constrained inverse covariance estimator (DICE) which achieves this information theoretic lower bound. They also propose another algorithm called sparse least-squares inverse covariance estimator (SLICE) which uses mixed integer quadratic programming, making it more efficient, but the sample complexity of SLICE is roughly $1/\kappa^2$ higher than the information theoretic lower bound.

1.5.8 Ill-Conditioned GGMs

CLIME or Graphical Lasso is only suitable when the precision matrix is well-conditioned. Kelnner et al. [28] considers the problem of estimating an ill-conditioned precision matrix in some important class of GGMS. They give fixed polynomial-time algorithms for learning *attractive GGMs* and *walk-summmable GGMs* with a logarithmic number of samples. Attractive GGMs are GGMs in which the off-diagonal entries of Θ are non-positive. This means that all partial correlations are non-negative. These are often used in practice, for example

in phylogenetic applications, observed variables are often positively dependent because of shared ancestry (see Zwiernik [60]); also in finance where using a latent global market variable leads to positive dependence (see Müller and Scarsini [39]). Kelner et al. [28] introduces an algorithm called GREEDY-AND-PRUNE which has sample complexity $\log(1/\kappa)$ times higher than the information theoretic lower bound but is more efficient than other methods. Walk-summable GGMs are defined as follows, see Malioutov et al. [35] for more details. A walk of length $l \geq 0$ in a graph $G = (V(G), E(G))$ is a sequence $w = (w_0, w_1, \dots, w_l)$ of nodes $w_k \in V(G)$ such that each step of the walk, say (w_k, w_{k+1}) , corresponds to an edge of the graph $\{w_k, w_{k+1}\} \in E(G)$. Walks may visit nodes and cross edges multiple times. We let $l(w)$ denote the length of walk w . We define the weight of a walk to be the product of the edge weights along the walk:

$$\phi(w) = \prod_{k=1}^{l(w)} r_{w_{k-1}, w_k}.$$

We also allow zero-length “self” walks $w = (v)$ at each node v for which we define $\phi(w) = 1$. The connection between these walks and Gaussian inference can be seen as follows. We decompose the covariance matrix as

$$\Sigma = \Theta^{-1} = (\mathbf{I} - \mathbf{R})^{-1} = \sum_{k=0}^{\infty} \mathbf{R}^k,$$

for $\rho(\mathbf{R}) < 1$. We have assumed that the model is normalized by rescaling variables so that $\Theta[i, i] = 1$ for all i . Then $\mathbf{R} = \mathbf{I} - \Theta$ has zero diagonal and the off-diagonal elements are equal to the partial correlation coefficients r_{ij}

$$r_{ij} := \frac{\text{Cov}(x_i, x_j | x_{V \setminus \{i,j\}})}{\sqrt{\text{Var}(x_i | x_{V \setminus \{i,j\}})} \sqrt{\text{Var}(x_j | x_{V \setminus \{i,j\}})}}.$$

The (i, j) -th entry of \mathbf{R}^l is sum over weights of paths of length l that go from i to j . A GGM is called walk-summable if for all i, j the sum of $|\phi(w)|$ over all walks from i to j is finite almost surely. For learning walk-summable GGMs [28] introduces an algorithm called HYBRIDMB which has sample complexity $1/\kappa^2$ times the information theoretic lower bound.

1.5.9 Bayesian Approaches

Bayesian methods have also been utilized for estimation of precision matrices. Banerjee and Ghosal [5] considers a prior distribution on the off-diagonal entries of the precision matrix which put a mixture of a point mass at zero and certain absolutely continuous distribution. They establish posterior consistency of the resulting estimator. Posterior consistency was established for class of banded precision matrix in Banerjee and Ghosal [4]. In Shi et al. [46] the authors consider the situation where the observation from the Graphical model are tampered with Gaussian measurement errors.

1.6 Our Contributions and Future Directions

In this work, we contribute a novel estimator for the weighted network from samples of a GFF on the network, based on certain Fourier analytic properties of the Gaussian distribution.

In this pursuit, our approach exploits complex-valued statistics constructed from observed data, that are of interest in their own right.

To wit, we begin with the observation that the logarithm of a probability density in a Gaussian Free Field is essentially a quadratic form of the network Laplacian (up to an additive constant), and is thus of interest in learning the Laplacian via its quadratic forms. Fundamentally, our approach is underpinned by the observation that the standard Gaussian density is essentially a fixed point of the Fourier transform, and for a general covariance matrix Σ , the Fourier transform entails a mapping $\Sigma \mapsto \Sigma^{-1}$ on the exponent of the Gaussian density. Thus, two successive applications of the Fourier transformation acts like an involution on a Gaussian density, up to a normalization constant. While taking the Fourier transform via numerical integration can be challenging computationally, we tackle this via a stochastic approach, by averaging against an independent Gaussian with a suitable dispersion. Our test statistic, which is complex-valued, is conveniently bounded in absolute value by 1, thereby allowing for superior concentration of measure effects.

Complex-valued statistical observables have certain salutary properties, which can be of interest from theoretical perspectives. In particular, they embody a phase, which can lead to stronger cancellation effects due to the destructive interference of phases. Such statistics, however, have not been exploited to their full potential, and literature on their application is quite limited. A recent instance in the literature is accorded by Belomestny et al. [7], where the authors use complex-valued test statistics in order to perform deconvolution in the context of covariance matrix estimation.

The approach proposed in the present work is conceptually simple and computationally light, in addition to having highly tractable analytical properties. Most of the known techniques for precision matrix estimation for GGMs are known to be of limited effectiveness when the GGM is ill-conditioned, such as the GFF (with a small mass parameter). Furthermore, the known techniques for graph recovery from GGMs, especially in the ill-conditioned setting (see, e.g., Kelner et al. [29]) are often combinatorial in nature and are more suited for the setting of unweighted graphs. Much of the existing literature is also geared towards the learning of sparse graphs (in the context of the high-dimensional tradeoff between system size and data availability), and involve computationally intensive optimization procedures. However, it may be noted that a network can be *low dimensional* without being sparse – this is significant vis-a-vis current interest in generative models, where a dense network can be generated from generative model with only a few parameters. A classic case in point is that of the Erdos-Renyi random graph in the dense regime, which is characterised by a single parameter, namely the edge connection probability; another instance on similar lines is provided by a stochastic block model.

Our approach addresses many of these issues with a simple and easy-to-compute estimator. We demonstrate the effectiveness of our estimator with concrete recovery guarantees and bounds on the required sample complexity. In particular, we show that the proposed statistic achieves the parametric rate of estimation for fixed network size. In the setting of networks growing with sample size, our results show that for Erdos–Renyi random graphs $G(d, p)$ above the connectivity threshold, network recovery takes place with high probability as soon as the sample size n satisfies $n \gg d^4 \log d \cdot p^{-2}$.

We believe that the present work inaugurates the study of complex-valued statistics and techniques inspired by Gaussian Fourier analysis in the context of GGMs, and more generally, Gaussian random fields with a geometric structure. A direction of particular interest would be to augment our simple approach with additional ingredients so as to account for structured network models, a case in point being that of sparsity. Further improvisations and modifications of our relatively straightforward approach to provide efficient learning in

wider classes of networks and enhanced rates in specific structural scenarios provide natural avenues for further investigation.

2 Learning Networks from Gaussian Free Fields

In this section, we will lay out our approach to estimation of the weighted network (equivalently, its Laplacian) based on Gaussian Fourier analysis. To fix notations, we set $\Sigma = (\mathbf{L} + \mu\mathbf{I})^{-1}$ and \mathbf{L} as $d \times d$ matrices and $\mu > 0$. Let

$$\lambda_1 := \lambda_{\max}(\mathbf{L}).$$

For $\eta > 0$, we define

$$\mathbf{L}^{(\eta)} := (\Sigma + \eta^{-1}\mathbf{I})^{-1}. \quad (4)$$

Our estimation procedure will comprise of two steps. We will first estimate $\mathbf{L}^{(\eta)}$, and then from this estimate of $\mathbf{L}^{(\eta)}$ we will construct an estimator of Σ^{-1} .

2.1 Estimation of $\mathbf{L}^{(\eta)}$

Our approach to estimation of $\mathbf{L}^{(\eta)}$ is based on the Fourier analytic properties of the Gaussian distribution. In particular, up to constants, the standard Gaussian density in d dimensions is a fixed point of the Fourier transform. For a general covariance matrix Σ , the Fourier transform induces a mapping $\Sigma \mapsto \Sigma^{-1}$ on the exponent of the Gaussian density. Thus, two successive applications of the Fourier transformation acts like an involution on a Gaussian density, up to a normalization constant.

Therefore, if there is a simple way to mimic the Fourier transform of an underlying Gaussian based on data generated from that distribution, then a twofold application of the Fourier transform (followed by a logarithmic transformation) would approximately result in a quadratic form in the precision matrix. Obtaining the precision matrix from this quadratic form would then be a rather simple matter.

In expectation, taking the Fourier transform of a Gaussian density on the basis of random samples from it is relatively canonical: we simply consider the plane wave corresponding to the random variable (or more precisely, its empirical version). To wit, if \mathbf{W} is a d -dimensional Gaussian random vector, then $\mathbb{E}\left(\exp(i\langle \xi, \mathbf{W} \rangle)\right)$ is the Fourier transform of the density of \mathbf{W} , evaluated at $\xi \in \mathbb{R}^d$. In practice, we do not work at the level of expectations, but based on a large set of samples drawn from \mathbf{W} ; thus the act of taking expectation is substituted canonically with averaging over this sample. Thus, we are performing a *stochastic* version of a Fourier transformation, based on observed samples from a distribution. In this vein, it is convenient that our test statistic (which is notably complex-valued) is conveniently bounded in absolute value by 1, thereby allowing for strong concentration of measure effects.

A second application of the Fourier transform would nominally entail another integral (in the variable ξ) against the complex harmonic $e^{i\langle \xi, \mathbf{t} \rangle}$. However, numerically integrating statistical estimates, such as those obtained from the first round of stochastic Fourier transform above, can be rather challenging, with attendant numerical instability effects. We tackle this issue by replacing the Lebesgue measure in this second integral by a Gaussian density with a suitable variance η ; the intuitive idea being that the true integral can be seen as a limit when $\eta \rightarrow \infty$. This introduces the additional parameter η into our estimation procedure, but this is

not too difficult to eliminate, and is the focus of the subsequent step of the process, discussed in the next section.

In this work, we combine the two steps above in one stroke, by considering the estimator in (7), which in turn is motivated by the expectation-level quantity φ in (5) and (6). In fact, an intuitively clarifying interpretation to (5) would be to first take the expectation with respect to the random variable \mathbf{X} (corresponding to the first round of Fourier transform discussed above), followed by expectation with respect to the variable \mathbf{Y} (corresponding to the second round of Fourier transformation in our earlier discussion).

Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined as

$$\varphi(\mathbf{t}) := \mathbb{E}(\exp(i\langle \mathbf{Y}, \mathbf{X} + \mathbf{t} \rangle)) \tag{5}$$

where $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \eta\mathbf{I})$, and $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the usual inner product

$$\langle \mathbf{y}, \mathbf{x} \rangle := \sum_{j=1}^d y_j x_j.$$

An alternative expression of $\varphi(\mathbf{t})$ is the following (see Lemma 5.1)

$$\varphi(\mathbf{t}) = \det\left(\frac{1}{\eta}\mathbf{L}^{(\eta)}\right)^{1/2} \cdot \exp\left(-\frac{1}{2}\langle \mathbf{t}, \mathbf{L}^{(\eta)}\mathbf{t} \rangle\right). \tag{6}$$

Estimating $\varphi(\mathbf{t})$ for well-chosen values of $\mathbf{t} \in \mathbb{R}^d$, and suitably aggregating these estimates, we will obtain an estimate of $\mathbf{L}^{(\eta)}$.

The sample version of the definition (5) of $\varphi(\mathbf{t})$ naturally suggests the following unbiased estimator:

$$\varphi_n(\mathbf{t}) := \frac{1}{n} \sum_{k=1}^n \exp(i\langle \mathbf{Y}_k, \mathbf{X}_k + \mathbf{t} \rangle), \tag{7}$$

where $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are i.i.d. random vectors in \mathbb{R}^d which are independent of the \mathbf{X}_j s with the common distribution $\mathcal{N}(\mathbf{0}, \eta\mathbf{I})$. Let $\mathbf{e}_1, \dots, \mathbf{e}_d$ denote the standard basis of \mathbb{R}^d .

Let $l_{ij}^{(\eta)}$ denote the (i, j) -entry of $\mathbf{L}^{(\eta)}$. A direct computation gives

$$l_{ij}^{(\eta)} = -2 \log \left| \varphi\left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}\right) \right| + \log |\varphi(\mathbf{e}_i)| + \log |\varphi(\mathbf{e}_j)| \text{ for } i \neq j,$$

$$l_{ii}^{(\eta)} = -2 \log |\varphi(\mathbf{e}_i)| + 2 \log |\varphi(\mathbf{0})|.$$

Due to this observation, we propose the following estimator of $\mathbf{L}^{(\eta)} = [l_{ij}^{(\eta)}]_{1 \leq i, j \leq d}$:

$$\widehat{\mathbf{L}}^{(\eta)} := \left[\widehat{l}_{ij}^{(\eta)} \right]_{1 \leq i, j \leq d}, \tag{8}$$

where

$$\widehat{l}_{ij}^{(\eta)} := -2 \log \left| \varphi_n\left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}\right) \right| + \log |\varphi_n(\mathbf{e}_i)| + \log |\varphi_n(\mathbf{e}_j)| \text{ for } i \neq j, \text{ and} \tag{9}$$

$$\widehat{l}_{ii}^{(\eta)} := -2 \log |\varphi_n(\mathbf{e}_i)| + 2 \log |\varphi_n(\mathbf{0})|. \tag{10}$$

2.2 Estimation of Σ^{-1} from $L^{(\eta)}$

The quantity $L^{(\eta)}$ estimated in the previous section involves the parameter η that is an artefact of our procedure. In this section our goal is to put forward a principled way of eliminating the parameter η and obtain an estimate of Σ^{-1} , which is our object of interest.

A vanilla approach to obtaining Σ^{-1} from $L^{(\eta)}$, in light of its defining equation (4), is to observe that Σ^{-1} and $L^{(\eta)}$ commute, and therefore spectral inversion is a possibility. However, direct inversion at the level of eigenvalues can lead to numerical instabilities, and instead, we propose the following approach, based on the so-called *Woodbury's identity*.

We first write down an identity relating Σ^{-1} and $L^{(\eta)}$. To this end, we recall the Woodbury matrix identity [53]:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{11}$$

With $A = \eta^{-1}I$, $C = \Sigma$, $U = I$ and $V = I$, we get

$$L^{(\eta)} = \eta I - \eta I(\Sigma^{-1} + \eta I)^{-1}\eta I = \eta I - \eta^2(\Sigma^{-1} + \eta I)^{-1}.$$

Thus

$$\Sigma^{-1} = \eta^2(\eta I - L^{(\eta)})^{-1} - \eta I. \tag{12}$$

Thus from an estimator $\widehat{L}^{(\eta)}$ of $L^{(\eta)}$, one can construct a plug-in estimator of Σ^{-1} :

$$\widehat{\Sigma}^{-1} = \eta^2(\eta I - \widehat{L}^{(\eta)})^{-1} - \eta I. \tag{13}$$

2.3 A Vanilla Spectral Estimator

In [7], the authors introduced the following estimator for the covariance matrix. For $\psi_n(\mathbf{u})$ the empirical version of the characteristic function of the Gaussian field, given by

$$\psi_n(\mathbf{u}) := \frac{1}{n} \sum_{j=1}^n \exp(i \langle \mathbf{u}, \mathbf{X}_j \rangle),$$

and a suitably chosen large parameter $U > 0$, we define

$$(\widehat{\Sigma}_{\text{BMT}})_{ii} := -\frac{2}{U^2} \Re(\log \psi_n(U\mathbf{e}_i)) \text{ and} \tag{14}$$

$$(\widehat{\Sigma}_{\text{BMT}})_{ij} := -\frac{2}{U^2} \Re\left(\log \psi_n\left(U \cdot \frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}\right)\right) - \frac{1}{2} ((\widehat{\Sigma}_{\text{BMT}})_{ii} + (\widehat{\Sigma}_{\text{BMT}})_{jj}) \text{ for } i \neq j. \tag{15}$$

A naive approach to estimating Σ^{-1} would be to invert $\widehat{\Sigma}_{\text{BMT}}$ directly. Conceptually, a principal point of divergence in which this estimator differs from our approach is that, instead of directly computing the matrix inverse Σ^{-1} , we apply another round of Gaussian Fourier analysis and access Σ^{-1} indirectly via that route.

Since both approaches involve Gaussian Fourier analytic or spectral ideas and complex valued statistics, it would be of interest to compare the two techniques. In particular, such comparison will clarify whether accessing the Laplacian indirectly via Gaussian Fourier analysis brings in any statistical benefits.

In summary, our analysis appears to corroborate the fact that the application of Gaussian Fourier analysis to access the Laplacian indirectly, as in our approach, brings in significant

statistical benefits. As such, this makes the case for wider investigation of similar ideas in the study of Gaussian random fields in general and Gaussian Graphical Models in particular.

3 Theoretical Guarantees

In this section, we lay out theoretical guarantees that demonstrate the effectiveness of our method.

3.1 Estimation Rates via Concentration Bounds

We first state a result showing how the error in estimating $\mathbf{L}^{(\eta)}$ influences the estimation error of $\widehat{\Sigma}^{-1}$.

Theorem 3.1 *Suppose that*

$$\frac{\lambda_1 + \mu + \eta}{\eta^2} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_2 < 1.$$

Then we have

$$\frac{1}{d} \|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|_F \leq \frac{(\lambda_1 + \mu + \eta)^2}{\eta^4 \left(1 - \frac{\lambda_1 + \mu + \eta}{\eta^2} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_2\right)} \frac{1}{d} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_F. \tag{16}$$

Theorem 3.1 naturally leads to the question of concentration bounds for $\widehat{\mathbf{L}}^{(\eta)}$, which we take up in the next section.

3.2 Concentration of $\widehat{\mathbf{L}}^{(\eta)}$

We begin with a concentration bound for φ_n .

Proposition 3.2 (Concentration of φ_n) *For any $x > 0$ and $\mathbf{t} \in \mathbb{R}^d$, we have*

$$\mathbb{P}(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq x) \leq 4 \exp\left(-\frac{3nx^2}{24 + 8x}\right).$$

In particular, for $x \in (0, 1]$ and $\mathbf{t} \in \mathbb{R}^d$, we have

$$\mathbb{P}(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq x) \leq 4 \exp\left(-\frac{3}{32}nx^2\right).$$

Let

$$c_\eta := \det\left(\frac{1}{\eta}\mathbf{L}^{(\eta)}\right)^{1/2}, \quad c_*(\eta) := \frac{1}{2}c_\eta \exp\left(-\frac{1}{2}\|\mathbf{L}^{(\eta)}\|_2^2\right), \tag{17}$$

and

$$S_n(\mathbf{t}) := \log |\varphi_n(\mathbf{t})| - \log |\varphi(\mathbf{t})|. \tag{18}$$

This allows us to state a concentration bound for S_n .

Proposition 3.3 (Concentration of S_n) For any $\mathbf{t} \in \mathbb{R}^d, \|\mathbf{t}\| \leq 1$ and $x \in (0, 1]$, we have

$$\mathbb{P}(|S_n(\mathbf{t})| \geq x) \leq 3\mathbb{P}(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq c_*(\eta)x).$$

Therefore, for any $\mathbf{t} \in \mathbb{R}^d, \|\mathbf{t}\| \leq 1$ and $x \in (0, 1]$, we have

$$\mathbb{P}(|S_n(\mathbf{t})| \geq x) \leq C_1 \exp(-C_2 \cdot nc_*(\eta)^2 x^2),$$

for some universal positive constants C_1, C_2 .

Concentration of $\varphi_n(\mathbf{t})$ around $\varphi(\mathbf{t})$ yields concentration of $\widehat{\mathbf{L}}^{(\eta)}$ around $\mathbf{L}^{(\eta)}$ as shown by the following lemma.

Lemma 3.4 (Concentration of φ_n implies concentration of $\widehat{\mathbf{L}}^{(\eta)}$) We have

$$\begin{aligned} l_{ii}^{(\eta)} - \widehat{l}_{ii}^{(\eta)} &= 2S_n(\mathbf{e}_i) - 2S_n(\mathbf{0}) \\ l_{ij}^{(\eta)} - \widehat{l}_{ij}^{(\eta)} &= 2S_n\left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}\right) - S_n(\mathbf{e}_i) - S_n(\mathbf{e}_j) \text{ for } i \neq j. \end{aligned}$$

Finally, we are ready to state

Theorem 3.5 (Concentration of $\widehat{\mathbf{L}}^{(\eta)}$) For $x \in (0, 1]$ we have

$$\mathbb{P}\left(\frac{1}{d}\|\mathbf{L}^{(\eta)} - \widehat{\mathbf{L}}^{(\eta)}\|_F \geq x\right) \leq d^2 \cdot C_1 \cdot \exp(-C_2 nc_*(\eta)^2 x^2),$$

for some universal positive constants C_1, C_2 . In other words, with probability at least $1 - d^{-c}$, we have

$$\frac{1}{d}\|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_F = O\left(\frac{1}{c_*(\eta)}\sqrt{\frac{\log d}{n}}\right).$$

Note that

$$c_*(\eta) = \frac{1}{2}c_\eta \exp\left(-\frac{1}{2}\|\mathbf{L}^{(\eta)}\|_2^2\right)$$

and

$$\|\mathbf{L}^{(\eta)}\|_2 = \frac{\eta}{1 + \eta\lambda_{\min}(\boldsymbol{\Sigma})}.$$

Thus $c_*(\eta) \asymp c_\eta$. But

$$\begin{aligned} c_\eta &= \left(\prod_{j=1}^d \frac{\lambda_j + \mu}{\lambda_j + \mu + \eta}\right)^{1/2} = \left(\prod_{j=1}^d \left(1 + \frac{\eta}{\lambda_j + \mu}\right)^{-1}\right)^{1/2} \\ &= \exp\left(-\frac{1}{2}\sum_{j=1}^d \log\left(1 + \frac{\eta}{\lambda_j + \mu}\right)\right). \end{aligned}$$

3.3 An Explicit Guarantee on the Estimation Rate

It remains to combine the main results of the last two sections to provide an explicit guarantee on estimation rates.

Theorem 3.6 *Suppose that*

$$\frac{\lambda_1 + \mu + \eta}{\eta^2} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_2 < 1.$$

Then, with probability $\geq 1 - d^{-c}$, we have

$$\frac{1}{d} \|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|_F \leq C \cdot \frac{(\lambda_1 + \mu + \eta)^2}{\eta^4 \left(1 - \frac{\lambda_1 + \mu + \eta}{\eta^2} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_2\right)} \cdot \frac{1}{c_*(\eta)} \sqrt{\frac{\log d}{n}} \tag{19}$$

for some positive constant C .

4 Generative Models: Erdős–Rényi Random Graphs

In this section, we investigate the behaviour of our approach in the setting of some common generative models of random graphs, and compare with the vanilla spectral estimator based on Belomestny et al. [7]. In particular, we will focus on the setup where the base graph $G(d, p)$ is generated from the Erdős–Rényi model with edge probability

$$p = \Omega\left(\frac{\log d}{d}\right).$$

We note in passing that the above regime of connection probability ensures that a graph generated according to this model is connected with high probability.

4.1 Estimating Erdős–Rényi Random Graphs via Our Approach

Herein we carry out a performance analysis of our approach for the Erdős–Rényi random graph model.

To state our main result of this section, we denote the average degree of the graph by

$$\Delta_{\text{avg}} := (d - 1)p.$$

We may now state

Theorem 4.1 *If we choose $\eta = \Theta(p)$ and take*

$$n > \frac{d^6 \log d}{\Delta_{\text{avg}}^2},$$

then we have with probability at least $1 - d^{-c}$ that

$$\frac{1}{d} \|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|_F \leq C \sqrt{\frac{d^4 \log d}{p^4 n}},$$

for some absolute constants $c, C > 0$.

Remark 4.2 It may be noted that in the dense regime of the Erdős–Rényi random graph model, i.e. with the connection probability $p = \Theta(1)$ (as $d \rightarrow \infty$), Theorem 4.1 implies a sample complexity of order $d^4 \log d$.

Remark 4.3 More generally, one can prove a similar result for inhomogeneous Erdős–Rényi random graphs with edge probability matrix P . There the role of the average degree Δ_{avg} would be replaced by the maximum expected degree $\Delta_{\text{max}} := \max_i \sum_j P_{ij}$.

4.2 Comparison to the Vanilla Spectral Estimator

In this section, we will provide a comparison of the theoretical guarantees on our approach vis-a-vis the vanilla spectral estimator motivated by Belomestny et al. [7]. The application of the latter estimator requires a choice of the parameter U . We undertake an analysis of two such possible choices—one following the recommendation of Belomestny et al. [7], and another following an improvisation tailored to our specific setting.

For both choices (c.f. (25), (26)) of parameter U , it appears from the analysis in the two subsequent sections that for dense Erdős–Rényi random graphs (i.e. $p = \Theta(1)$), our approach has a better sample complexity guarantee for ill-conditioned GFFs (i.e., the mass parameter μ being small); refer to Theorem 4.1 and Remark 4.2.

For our purpose, we state here a simplified version of Theorem 1 of [7]. Note that in [7], the observations are $\mathbf{Y}_i = \mathbf{X}_i + \boldsymbol{\varepsilon}_i, i = 1, \dots, n$, where $\boldsymbol{\varepsilon}_i$ s are i.i.d. noises independent of \mathbf{X}_i s. In our setting, the noises $\boldsymbol{\varepsilon}_i = 0$.

Theorem 4.4 (Theorem 1 of [7]) *Assume that $\|\boldsymbol{\Sigma}\|_2 \leq R$. Let $\gamma > \sqrt{2}$ and $U \geq 1$ satisfy*

$$8\gamma \sqrt{\frac{\log(ed)}{n}} < e^{-RU^2}. \tag{20}$$

Set

$$\tau(U) := 6\gamma \frac{e^{RU^2}}{U^2} \left(\frac{\log(ed)}{n}\right)^{1/2}. \tag{21}$$

Then for any $\tau \geq \tau(U)$,

$$\mathbb{P}(\|\widehat{\boldsymbol{\Sigma}}_{\text{BMT}} - \boldsymbol{\Sigma}\|_\infty < \tau) \geq 1 - 12e^{-\gamma^2 d^{2-\gamma^2}}.$$

To undertake the analysis for the vanilla estimator, we also need the following preparatory result.

Proposition 4.5 *Let $\widehat{\boldsymbol{\Sigma}}$ denote any estimator of $\boldsymbol{\Sigma}$. Assume that $s_{\min}(\boldsymbol{\Sigma}) > \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$, where $s_{\min}(\mathbf{M})$ denotes the smallest singular value of \mathbf{M} . Then*

$$\frac{1}{d} \|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F \leq \frac{\frac{1}{d} \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F}{s_{\min}(\boldsymbol{\Sigma})(s_{\min}(\boldsymbol{\Sigma}) - \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2)}.$$

Let us now analyse the vanilla spectral estimator $\widehat{\boldsymbol{\Sigma}}_{\text{BMT}}^{-1}$ with two specific choices of parameter U : the *canonical choice* (recommended by Belomestny et al. in [7]) and an *improvised choice* (tailored to our specific setting). In [7], the authors suggested the *canonical choice* for U should take the form

$$U = c_0 R^{-1/2} \sqrt{\log\left(\frac{n}{\log(ed)}\right)} \tag{22}$$

for some sufficiently small positive constant c_0 . On the other hand, in the light of Theorem 4.4 (a simplified version of Theorem 1 in [7] which is tailored to our setting), if we fix the parameter $\gamma > \sqrt{2}$ then the condition (20) is equivalent to

$$RU^2 < \frac{1}{2} \log \left(\frac{n}{\log(ed)} \right) - \log(8\gamma). \tag{23}$$

The optimal choice for the parameter U should minimize the quantity $\tau(U)$ in (21). Thus, provided that n is large enough, the *improvised choice* for U should be

$$U = R^{-1/2}. \tag{24}$$

Theorem 4.6 (i) (*Canonical parameter choice*) If U is of the form (22), then for any $\alpha \in (0, 1/2)$, we have

$$\frac{1}{d} \|\widehat{\Sigma}_{\text{BMT}}^{-1} - \Sigma^{-1}\|_F = O\left(\frac{(\lambda_1 + \mu)^2}{\mu} \left(\frac{\log(ed)}{n}\right)^{1/2-\alpha}\right)$$

with probability at least $1 - Cd^{-c}$ for some constants $c, C > 0$, provided that

$$n \geq \left(\frac{d(\lambda_1 + \mu)}{\mu}\right)^{\frac{2}{1-2\alpha}} \log(ed).$$

(ii) (*Improvised parameter choice*) If U is of the form (24), then we have

$$\frac{1}{d} \|\widehat{\Sigma}_{\text{BMT}}^{-1} - \Sigma^{-1}\|_F = O\left(\frac{(\lambda_1 + \mu)^2}{\mu} \left(\frac{\log(ed)}{n}\right)^{1/2}\right)$$

with probability at least $1 - Cd^{-c}$ for some constants $c, C > 0$, provided that

$$n \geq \left(\frac{d(\lambda_1 + \mu)}{\mu}\right)^2 \log(ed).$$

When \mathbf{L} is the Laplacian of an $\text{ER}(d, p)$ random graph, we have $\lambda_1 = dp + O(\sqrt{dp \log d})$ with high probability. Applying Theorem 4.6 to this special case gives

Corollary 4.7 (i) (*Canonical parameter choice*) If U is of the form (22), then for any $\alpha \in (0, 1/2)$, we have

$$\frac{1}{d} \|\widehat{\Sigma}_{\text{BMT}}^{-1} - \Sigma^{-1}\|_F = O\left(\frac{(dp)^2}{\mu} \left(\frac{\log(ed)}{n}\right)^{1/2-\alpha}\right)$$

with probability at least $1 - Cd^{-c}$ for some constants $c, C > 0$, provided that

$$n \geq \left(\frac{d^2 p}{\mu}\right)^{\frac{2}{1-2\alpha}} \log(ed). \tag{25}$$

(ii) (*Improvised parameter choice*) If U is of the form (24), then we have

$$\frac{1}{d} \|\widehat{\Sigma}_{\text{BMT}}^{-1} - \Sigma^{-1}\|_F = O\left(\frac{(dp)^2}{\mu} \left(\frac{\log(ed)}{n}\right)^{1/2}\right)$$

with probability at least $1 - Cd^{-c}$ for some constants $c, C > 0$, provided that

$$n \geq \left(\frac{d^2 p}{\mu}\right)^2 \log(ed). \tag{26}$$

5 Detailed Proofs of Theoretical Results

In this section, we provide detailed proofs of various theoretical results in the earlier sections of the paper.

5.1 On the Expectation of $\varphi_n(\mathbf{t})$

We begin with a Lemma that deals with the expectation of $\varphi_n(\mathbf{t})$.

Lemma 5.1 (Expectation of $\varphi_n(\mathbf{t})$)

$$\varphi(\mathbf{t}) = \det\left(\frac{1}{\eta}\mathbf{L}^{(\eta)}\right)^{1/2} \cdot \exp\left(-\frac{1}{2}\langle\mathbf{t}, \mathbf{L}^{(\eta)}\mathbf{t}\rangle\right).$$

Proof We observe that

$$\begin{aligned} \varphi(\mathbf{t}) &= \mathbb{E}\left[\exp\left(i\langle\mathbf{Y}_1, \mathbf{X}_1 + \mathbf{t}\rangle\right)\right] \\ &= \mathbb{E}_{\mathbf{Y}_1}\left[\exp\left(i\langle\mathbf{Y}_1, \mathbf{t}\rangle\right)\mathbb{E}_{\mathbf{X}_1}\left[\exp\left(i\langle\mathbf{Y}_1, \mathbf{X}_1\rangle\right)\right]\right] \\ &= \mathbb{E}_{\mathbf{Y}_1}\left[\exp\left(i\langle\mathbf{Y}_1, \mathbf{t}\rangle\right)\exp\left(-\frac{1}{2}\langle\mathbf{Y}_1, \boldsymbol{\Sigma}\mathbf{Y}_1\rangle\right)\right] \\ &= \int_{\mathbb{R}^d} \frac{1}{\sqrt{\det(2\pi\eta\mathbf{I})}}\exp\left(i\langle\mathbf{y}, \mathbf{t}\rangle\right)\exp\left(-\frac{1}{2}\langle\mathbf{y}, \boldsymbol{\Sigma}\mathbf{y}\rangle\right)\exp\left(-\frac{1}{2}\langle\mathbf{y}, \eta^{-1}\mathbf{y}\rangle\right) d\mathbf{y} \\ &= \int_{\mathbb{R}^d} \frac{1}{\sqrt{\det(2\pi\eta\mathbf{I})}}\exp\left(i\langle\mathbf{y}, \mathbf{t}\rangle\right)\exp\left(-\frac{1}{2}\langle\mathbf{y}, (\boldsymbol{\Sigma} + \eta^{-1}\mathbf{I})\mathbf{y}\rangle\right) d\mathbf{y} \\ &= \frac{1}{\sqrt{\det(\mathbf{I} + \eta\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}\langle\mathbf{t}, \mathbf{L}^{(\eta)}\mathbf{t}\rangle\right) \\ &= c_\eta\exp\left(-\frac{1}{2}\langle\mathbf{t}, \mathbf{L}^{(\eta)}\mathbf{t}\rangle\right), \end{aligned}$$

where

$$c_\eta := \det(\mathbf{I} + \eta\boldsymbol{\Sigma})^{-1/2} = \det\left(\frac{1}{\eta}\mathbf{L}^{(\eta)}\right)^{1/2}. \tag{27}$$

□

5.2 Proof of Lemma 3.4

We continue with the proof of Lemma 3.4.

Proof of Lemma 3.4 Observe that

$$\begin{aligned} \frac{1}{2}\langle\mathbf{t}, \mathbf{L}^{(\eta)}\mathbf{t}\rangle &= -\log\varphi(\mathbf{t}) + \log\varphi(\mathbf{0}) \\ &= -\log\varphi_n(\mathbf{t}) + \left(\log\varphi_n(\mathbf{t}) - \log\varphi(\mathbf{t})\right) + \log\varphi(\mathbf{0}). \end{aligned}$$

Taking real parts of both sides yields

$$\begin{aligned} \frac{1}{2} \langle \mathbf{t}, \mathbf{L}^{(n)} \mathbf{t} \rangle &= -\log |\varphi_n(\mathbf{t})| + \left(\log |\varphi_n(\mathbf{t})| - \log |\varphi(\mathbf{t})| \right) + \log |\varphi(\mathbf{0})| \\ &= -\log |\varphi_n(\mathbf{t})| + S(\mathbf{t}) + \log |\varphi(\mathbf{0})|. \end{aligned}$$

Let $l_{ij}^{(\eta)}$ denote the entry at i 'th row and j 'th column of $\mathbf{L}^{(n)}$. Since $\mathbf{L}^{(n)}$ is symmetric, we have

$$\langle \mathbf{e}_i, \mathbf{L}^{(n)} \mathbf{e}_i \rangle = l_{ii}^{(\eta)}, \quad \left\langle \frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}, \mathbf{L}^{(n)} \left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}} \right) \right\rangle = \frac{1}{2} l_{ii}^{(\eta)} + l_{ij}^{(\eta)} + \frac{1}{2} l_{jj}^{(\eta)} \quad \text{for } i \neq j.$$

Recall that

$$\begin{aligned} \widehat{l_{ii}^{(\eta)}} &= -2 \log |\varphi_n(\mathbf{e}_i)| + 2 \log |\varphi_n(\mathbf{0})| \\ \widehat{l_{ij}^{(\eta)}} &= -2 \log \left| \varphi_n \left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}} \right) \right| + \log |\varphi_n(\mathbf{e}_i)| + \log |\varphi_n(\mathbf{e}_j)| \quad \text{for } i \neq j. \end{aligned}$$

Thus,

$$\begin{aligned} l_{ii}^{(\eta)} - \widehat{l_{ii}^{(\eta)}} &= \langle \mathbf{e}_i, \mathbf{L}^{(n)} \mathbf{e}_i \rangle - (-2 \log |\varphi_n(\mathbf{e}_i)| + 2 \log |\varphi_n(\mathbf{0})|) \\ &= 2S_n(\mathbf{e}_i) - 2S_n(\mathbf{0}) \end{aligned}$$

and for $i \neq j$

$$\begin{aligned} l_{ij}^{(\eta)} - \widehat{l_{ij}^{(\eta)}} &= \left\langle \frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}, \mathbf{L}^{(n)} \left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}} \right) \right\rangle - \frac{1}{2} (l_{ii}^{(\eta)} + l_{jj}^{(\eta)}) - \widehat{l_{ij}^{(\eta)}} \\ &= 2S_n \left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}} \right) - S_n(\mathbf{e}_i) - S_n(\mathbf{e}_j). \end{aligned}$$

□

5.3 Proofs of Propositions 3.2 and 3.3

Our next item is the proof of Proposition 3.2.

Proof of Proposition 3.2 Note that

$$|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \leq |\Re(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))| + |\Im(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))|.$$

For $k \in \{1, \dots, n\}$, we define

$$\xi_k := \Re \left(e^{i \langle \mathbf{Y}_k, \mathbf{X}_k + \mathbf{t} \rangle} - \varphi(\mathbf{t}) \right).$$

Then we have i.i.d. real random variables ξ_1, \dots, ξ_n such that

$$\Re(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})) = \frac{1}{n} \sum_{k=1}^n \xi_k.$$

Observe that

$$\mathbb{E}(\xi_k) = 0, \quad |\xi_k| \leq 1 + c_\eta \leq 2, \quad \text{Var}(\xi_k) \leq 1 - |\varphi(\mathbf{t})|^2 \leq 1.$$

By the Bernstein’s inequality, we have for any $x > 0$

$$\mathbb{P}\left(\left|\Re(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))\right| \geq x\right) = \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n \xi_k\right| \geq x\right) \leq 2 \exp\left(-\frac{\frac{1}{2}nx^2}{1 + \frac{2}{3}x}\right).$$

Using similar bound for the imaginary part, we have

$$\begin{aligned} &\mathbb{P}(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq x) \\ &\leq \mathbb{P}\left(\left|\Re(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))\right| + \left|\Im(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))\right| \geq x\right) \\ &\leq \mathbb{P}\left(\left|\Re(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))\right| \geq \frac{x}{2}\right) + \mathbb{P}\left(\left|\Im(\varphi_n(\mathbf{t}) - \varphi(\mathbf{t}))\right| \geq \frac{x}{2}\right) \\ &\leq 4 \exp\left(-\frac{\frac{1}{8}nx^2}{1 + \frac{1}{3}x}\right) = 4 \exp\left(-\frac{3nx^2}{24 + 8x}\right). \end{aligned}$$

This completes the proof of Proposition 3.2. □

We continue on to the proof of Proposition 3.3.

Proof of Proposition 3.3 We have

$$S_n(\mathbf{t}) = \log\left|\frac{\varphi_n(\mathbf{t})}{\varphi(\mathbf{t})}\right| \leq \log\left(\left|\frac{\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})}{\varphi(\mathbf{t})}\right| + 1\right) \leq \left|\frac{\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})}{\varphi(\mathbf{t})}\right|,$$

which implies that

$$\mathbb{P}\left(S_n(\mathbf{t}) \geq x\right) \leq \mathbb{P}\left(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq x \cdot |\varphi(\mathbf{t})|\right) \text{ for any } x > 0.$$

On the event

$$\left\{|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \leq \frac{1}{2} |\varphi(\mathbf{t})|\right\}$$

we have

$$\begin{aligned} -S_n(\mathbf{t}) &= \log\left|\frac{\varphi(\mathbf{t})}{\varphi_n(\mathbf{t})}\right| \leq \log\left(\left|\frac{\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})}{\varphi_n(\mathbf{t})}\right| + 1\right) \\ &\leq \log\left(2\left|\frac{\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})}{\varphi(\mathbf{t})}\right| + 1\right) \leq 2\left|\frac{\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})}{\varphi(\mathbf{t})}\right|. \end{aligned}$$

Hence, for any $x > 0$,

$$\mathbb{P}\left(-S_n(\mathbf{t}) \geq x\right) \leq \mathbb{P}\left(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq \frac{x}{2} |\varphi(\mathbf{t})|\right) + \mathbb{P}\left(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| > \frac{1}{2} |\varphi(\mathbf{t})|\right).$$

In particular, if $x \in (0, 1]$ we deduce that

$$\mathbb{P}(|S_n(\mathbf{t})| \geq x) \leq 3 \cdot \mathbb{P}\left(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq \frac{x}{2} |\varphi(\mathbf{t})|\right).$$

On the other hand, for any $\mathbf{t} \in \mathbb{R}^d$ with $\|\mathbf{t}\| \leq 1$, one has

$$\left|\langle \mathbf{t}, \mathbf{L}^{(\eta)} \mathbf{t} \rangle\right| \leq \|\mathbf{L}^{(\eta)}\|_2^2.$$

This implies

$$|\varphi(\mathbf{t})| = c_\eta \exp\left(-\frac{1}{2} \langle \mathbf{t}, \mathbf{L}^{(\eta)} \mathbf{t} \rangle\right) \leq c_\eta \exp\left(-\frac{1}{2} \|\mathbf{L}^{(\eta)}\|_2^2\right) = 2c_*(\eta).$$

Thus, for $x \in (0, 1]$ and $\|\mathbf{t}\| \leq 1$

$$\mathbb{P}(|S_n(\mathbf{t})| \geq x) \leq 3\mathbb{P}(|\varphi_n(\mathbf{t}) - \varphi(\mathbf{t})| \geq c_*(\eta)x).$$

□

5.4 Proofs of Theorems 3.1 and 3.5

We first tackle the proof of Theorem 3.5

Proof of Theorem 3.5 For $x \in (0, 1]$, we have

$$\begin{aligned} \mathbb{P}\left(\left|l_{ii}^{(\eta)} - \widehat{l}_{ii}^{(\eta)}\right| \geq x\right) &\leq \mathbb{P}\left(2|S_n(\mathbf{e}_i)| + 2|S_n(\mathbf{0})| \geq x\right) \\ &\leq \mathbb{P}\left(|S_n(\mathbf{e}_i)| \geq \frac{x}{4}\right) + \mathbb{P}\left(|S_n(\mathbf{0})| \geq \frac{x}{4}\right) \\ &\leq C_1 \cdot \exp\left(-C_2nc_*(\eta)^2x^2\right), \end{aligned}$$

for some universal constants $C_1, C_2 > 0$.

For $i \neq j$ and $x \in (0, 1]$, we have

$$\begin{aligned} \mathbb{P}\left(\left|l_{ij}^{(\eta)} - \widehat{l}_{ij}^{(\eta)}\right| \geq x\right) &\leq \mathbb{P}\left(2\left|S_n\left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}\right)\right| + |S_n(\mathbf{e}_i)| + |S_n(\mathbf{e}_j)| \geq x\right) \\ &\leq \mathbb{P}\left(\left|S_n\left(\frac{\mathbf{e}_i + \mathbf{e}_j}{\sqrt{2}}\right)\right| \geq \frac{x}{4}\right) + \mathbb{P}\left(|S_n(\mathbf{e}_i)| \geq \frac{x}{4}\right) + \mathbb{P}\left(|S_n(\mathbf{e}_j)| \geq \frac{x}{4}\right) \\ &\leq C_1 \cdot \exp\left(-C_2nc_*(\eta)^2x^2\right) \end{aligned}$$

for some universal constants $C_1, C_2 > 0$.

Therefore, for $x \in (0, 1]$

$$\mathbb{P}\left(\max_{i,j} \left|l_{ij}^{(\eta)} - \widehat{l}_{ij}^{(\eta)}\right| \geq x\right) \leq d^2 \cdot C_1 \cdot \exp\left(-C_2nc_*(\eta)^2x^2\right),$$

for some universal constants $C_1, C_2 > 0$. Note that,

$$\|\mathbf{L}^{(\eta)} - \widehat{\mathbf{L}}^{(\eta)}\|_F^2 = \sum_{i,j} \left|l_{ij}^{(\eta)} - \widehat{l}_{i,j}^{(\eta)}\right|^2 \leq d^2 \cdot \max_{i,j} \left|l_{ij}^{(\eta)} - \widehat{l}_{ij}^{(\eta)}\right|^2.$$

□

We are now ready to address the proof of Theorem 3.1.

Proof of Theorem 3.1 From (12) and (13), we have

$$\widehat{\Sigma}^{-1} - \Sigma^{-1} = \eta^2 \left((\eta\mathbf{I} - \widehat{\mathbf{L}}^{(\eta)})^{-1} - (\eta\mathbf{I} - \mathbf{L}^{(\eta)})^{-1} \right).$$

Writing $\mathbf{X} = \mathbf{L}^{(\eta)}$ and $\mathbf{X}' = \widehat{\mathbf{L}}^{(\eta)}$, we have

$$\begin{aligned} (\eta\mathbf{I} - \mathbf{X}')^{-1} &= (\eta\mathbf{I} - \mathbf{X} + \mathbf{X} - \mathbf{X}')^{-1} \\ &= (\eta\mathbf{I} - \mathbf{X})^{-1} - (\eta\mathbf{I} - \mathbf{X})^{-1} (\mathbf{X} - \mathbf{X}') (\mathbf{I} + (\eta\mathbf{I} - \mathbf{X})^{-1} (\mathbf{X} - \mathbf{X}'))^{-1} (\eta\mathbf{I} - \mathbf{X})^{-1}, \end{aligned}$$

where for the second equality above we have used the Woodbury identity with $\mathbf{A} = \eta\mathbf{I} - \mathbf{X}$, $\mathbf{C} = \mathbf{I}$, $\mathbf{U} = \mathbf{X} - \mathbf{X}'$ and $\mathbf{V} = \mathbf{I}$. Now, using the inequalities $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$ and $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$, we obtain

$$\begin{aligned} & \|(\eta\mathbf{I} - \mathbf{X}')^{-1} - (\eta\mathbf{I} - \mathbf{X})^{-1}\|_F \\ &= \|(\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}')(\mathbf{I} + (\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}'))^{-1}(\eta\mathbf{I} - \mathbf{X})^{-1}\|_F \\ &\leq \|(\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}')\|_F \|(\mathbf{I} + (\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}'))^{-1}(\eta\mathbf{I} - \mathbf{X})^{-1}\|_2 \\ &\leq \|(\eta\mathbf{I} - \mathbf{X})^{-1}\|_2 \|\mathbf{X} - \mathbf{X}'\|_F \|(\mathbf{I} + (\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}'))^{-1}\|_2 \|(\eta\mathbf{I} - \mathbf{X})^{-1}\|_2 \\ &\leq \|\mathbf{X} - \mathbf{X}'\|_F \|(\eta\mathbf{I} - \mathbf{X})^{-1}\|_2^2 \|(\mathbf{I} + (\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}'))^{-1}\|_2. \end{aligned}$$

Now observe that $\eta\mathbf{I} - \mathbf{X} = \eta\mathbf{I} - \mathbf{L}^{(\eta)} = \eta^2(\boldsymbol{\Sigma}^{-1} + \eta\mathbf{I})^{-1}$, which implies $(\eta\mathbf{I} - \mathbf{X})^{-1} = \eta^{-2}(\boldsymbol{\Sigma}^{-1} + \eta\mathbf{I})$. Hence

$$\|(\eta\mathbf{I} - \mathbf{X})^{-1}\|_2 = \eta^{-2}(\lambda_1 + \mu + \eta).$$

Let $\mathbf{E} := (\eta\mathbf{I} - \mathbf{X})^{-1}(\mathbf{X} - \mathbf{X}')$. Then

$$\|\mathbf{E}\|_2 \leq \eta^{-2}(\lambda_1 + \mu + \eta)\|\mathbf{X} - \mathbf{X}'\|_2.$$

As long as $\|\mathbf{E}\|_2 < 1$, we have

$$\|(\mathbf{I} + \mathbf{E})^{-1}\|_2 \leq \frac{1}{1 - \|\mathbf{E}\|_2}.$$

Putting everything together, we get

$$\begin{aligned} \frac{1}{d} \|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F &\leq \frac{(\lambda_1 + \mu + \eta)^2}{\eta^4(1 - \|\mathbf{E}\|_2)} \cdot \frac{1}{d} \cdot \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_F \\ &\leq \frac{(\lambda_1 + \mu + \eta)^2}{\eta^4\left(1 - \frac{\lambda_1 + \mu + \eta}{\eta^2} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_2\right)} \cdot \frac{1}{d} \cdot \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_F. \end{aligned}$$

This completes the proof. □

5.5 Proof of Proposition 4.5

We now establish Proposition 4.5, which is of importance for the main results in Sect. 4.

Proof of Proposition 4.5 The identity

$$\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} = \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}$$

gives

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F &= \|\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}\|_F \\ &\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}\|_2 \|(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1}\|_F \\ &\leq \|\widehat{\boldsymbol{\Sigma}}^{-1}\|_2 \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_F \|\boldsymbol{\Sigma}^{-1}\|_2 \\ &\leq \frac{\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F}{s_{\min}(\widehat{\boldsymbol{\Sigma}})s_{\min}(\boldsymbol{\Sigma})}. \end{aligned} \tag{28}$$

Now, by Weyl’s inequality,

$$|s_{\min}(\widehat{\Sigma}) - s_{\min}(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_2,$$

which yields the lower bound

$$s_{\min}(\widehat{\Sigma}) \geq s_{\min}(\Sigma) - \|\widehat{\Sigma} - \Sigma\|_2 > 0.$$

Plugging this into (28) we get the desired upper bound. □

5.6 Proofs of Theorems 4.1 and 4.6

We complete this section with the proofs of the Theorems in Sect. 4.

Proof of Theorem 4.1 For semi-dense Erdős–Rényi graphs, i.e.,

$$p = \Omega\left(\frac{\log d}{d}\right),$$

all the non-zero eigenvalues of $\mathbf{L} = \mathbf{D} - \mathbf{A}$ are $\Theta(dp)$, with probability at least $1 - d^{-c}$. This follows from Weyl’s inequality that

$$|\lambda_i - (d - 1)p| \leq \|(\mathbf{D} - \mathbf{A}) - ((d - 1)p\mathbf{I} - p(\mathbf{J} - \mathbf{I}))\|_2$$

for $i = 1, \dots, d - 1$, and the estimate

$$\begin{aligned} & \| \mathbf{D} - \mathbf{A} - ((d - 1)p\mathbf{I} - p(\mathbf{J} - \mathbf{I})) \|_2 \\ & \leq \| \mathbf{D} - (d - 1)p\mathbf{I} \|_2 + \| \mathbf{A} - p(\mathbf{J} - \mathbf{I}) \|_2 \\ & = O\left(\sqrt{dp \log d}\right), \end{aligned}$$

which holds with probability at least $1 - d^{-c}$.

Indeed,

$$\| \mathbf{D} - (d - 1)p\mathbf{I} \|_2 = \max_i |D_{ii} - (d - 1)p|.$$

Now D_{ii} follows Binomial $((d - 1), p)$, and hence $D_{ii} - (d - 1)p$ is zero mean Sub-Gaussian with parameter $\sigma^2 = O(dp)$, whence it follows (c.f. Vershynin [50] Chap. 2) that with probability at least $1 - d^{-c}$ we have

$$\max_i |D_{ii} - (d - 1)p| = O\left(\sqrt{dp \log d}\right).$$

On the other hand,

$$\| \mathbf{A} - p(\mathbf{J} - \mathbf{I}) \|_2 \leq O\left(\sqrt{dp}\right),$$

with probability at least $1 - d^{-c}$ (see, e.g., Theorem 5.2 in [31]).

Thus if $\eta = O(p)$, we have

$$c_\eta \geq \exp\left(-\frac{1}{2} \sum_{j=1}^d \frac{\eta}{\lambda_j + \mu}\right) \asymp \exp\left(-C' \frac{\eta}{p}\right) = \Theta(1).$$

Therefore, with probability at least $1 - d^{-c}$, we have

$$\frac{1}{d} \|\widehat{\mathbf{L}}^{(\eta)} - \mathbf{L}^{(\eta)}\|_F = O\left(\sqrt{\frac{\log d}{n}}\right).$$

Using (19), we obtain

$$\begin{aligned} & \frac{1}{d} \|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|_F \\ & \leq \frac{(dp)^2}{p^4} \frac{1}{1 - \frac{d\|\widehat{\mathbf{L}}^{(n)} - \mathbf{L}^{(n)}\|_2}{p}} \frac{1}{d} \|\widehat{\mathbf{L}}^{(n)} - \mathbf{L}^{(n)}\|_F \\ & = O\left(\sqrt{\frac{d^4 \log d}{p^4 n}}\right), \end{aligned}$$

provided

$$\frac{d\|\widehat{\mathbf{L}}^{(n)} - \mathbf{L}^{(n)}\|_2}{p} \leq \frac{d^2}{p} \frac{1}{d} \|\widehat{\mathbf{L}}^{(n)} - \mathbf{L}^{(n)}\|_F \leq \frac{d^2}{p} \sqrt{\frac{\log d}{n}} < 1.$$

Thus we need a sample complexity of

$$n > \frac{d^6 \log d}{\Delta_{\text{avg}}^2}$$

to attain an estimation error of $O(\sqrt{\frac{d^4 \log d}{p^4 n}})$. □

We finally complete the proof of Theorem 4.6.

Proof of Theorem 4.6 First, we note that in our set-up,

$$\|\Sigma\|_2 = \mu^{-1}, \quad s_{\min}(\Sigma) = (\lambda_1 + \mu)^{-1}.$$

(i) From Theorem 1 of [7], we get by choosing $R = \|\Sigma\|_2 = \mu^{-1}$ and $U = c_0 \sqrt{\frac{1}{R} \log(n/\log(ed))}$ (where $c_0 > 0$ is a sufficiently small constant) that for any $\alpha \in (0, 1/2)$,

$$\|\widehat{\Sigma}_{\text{BMT}} - \Sigma\|_\infty = O\left(\mu^{-1} \left(\frac{\log(ed)}{n}\right)^{1/2-\alpha}\right),$$

with probability at least $1 - Cd^{-c}$ for some $c, C > 0$. As $\|\cdot\|_2 \leq \|\cdot\|_F \leq d\|\cdot\|_\infty$, we get from Proposition 4.5 that

$$\frac{1}{d} \|\widehat{\Sigma}_{\text{BMT}}^{-1} - \Sigma^{-1}\|_F = O\left((\lambda_1 + \mu)^2 \mu^{-1} \left(\frac{\log(ed)}{n}\right)^{1/2-\alpha}\right),$$

provided

$$d\mu^{-1} \left(\frac{\log(ed)}{n}\right)^{1/2-\alpha} < c_* s_{\min}(\Sigma) = c_*(\lambda_1 + \mu)^{-1},$$

for a sufficiently small $c_* > 0$, i.e.

$$n \geq \left(d\mu^{-1}(\lambda_1 + \mu)\right)^{\frac{2}{1-2\alpha}} \log(ed),$$

as desired.

(ii) From Theorem 1 of [7], we get by choosing $R = \|\Sigma\|_2 = \mu^{-1}$ and $U = R^{-1/2}$ that

$$\|\widehat{\Sigma}_{\text{BMT}} - \Sigma\|_\infty = O\left(\mu^{-1} \left(\frac{\log(ed)}{n}\right)^{1/2}\right)$$

with probability at least $1 - Cd^{-c}$ for some $c, C > 0$. Using similar argument as part (i), we get the desired result. This completes the proof. \square

Acknowledgements S.G. was supported in part by the MOE Grants R-146-000-250-133, R-146-000-312-114 and MOE-T2EP20121-0013. S.S.M. was partially supported by an INSPIRE research Grant (DST/INSPIRE/04/2018/002193) from the Department of Science and Technology, Government of India and a Start-Up Grant from Indian Statistical Institute, Kolkata. H.S.T. was supported by the NUS Research Scholarship. We thank Satya Majumdar for helpful suggestions.

Data Availability This manuscript has no associated data.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

1. Anandkumar, A., Tan, V., Willsky, A.: High-dimensional Gaussian graphical model selection: walk summability and local separation criterion. *J. Mach. Learn. Res.* **13**, 07 (2011)
2. Anandkumar, A., Tan, V.Y., Huang, F., Willsky, A.S.: High-dimensional structure estimation in Ising models: local separation criterion. *Ann. Stat.* **40**, 1346–1375 (2012)
3. Banerjee, O., Ghaoui, L.: Model selection through sparse max likelihood estimation model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.* **9**, 08 (2007)
4. Banerjee, S., Ghosal, S.: Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8**(2), 2111–2137 (2014)
5. Banerjee, S., Ghosal, S.: Bayesian structure learning in graphical models. *J. Multivar. Anal.* **136**, 147–162 (2015)
6. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* **37**(4), 382–390 (2005). <https://doi.org/10.1038/ng1532>
7. Belomestny, D., Trabs, M., Tsybakov, A.B.: Sparse covariance matrix estimation in high-dimensional deconvolution. *Bernoulli* **25**(3), 8 (2019). <https://doi.org/10.3150/18-BEJ1040A>
8. Berestycki, N.: Introduction to the Gaussian free field and Liouville quantum gravity. Lecture notes, 2018–2019 (2015)
9. Berthet, Q., Rigollet, P., Srivastava, P.: Exact recovery in the ising blockmodel. *Ann. Stat.* **47**(4), 1805–1834 (2019)
10. Bhattacharya, B. B., Mukherjee, S.: Inference in ising models. (2018)
11. Bickel, P.J., Levina, E.: Regularized estimation of large covariance matrices. *Ann. Stat.* **36**(1), 199–227 (2008)
12. Bickel, P.J., Levina, E.: Covariance regularization by thresholding. *Ann. Stat.* **36**(6), 2577–2604 (2008)
13. Bresler, G.: Efficiently learning ising models on arbitrary graphs. In: Proceedings of the forty-seventh annual ACM symposium on Theory of computing, pp. 771–782 (2015)
14. Cai, T.T., Zhang, C.-H., Zhou, H.H.: Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.* **38**(4), 2118–2144 (2010)
15. Cai, T.T., Li, H., Liu, W., Xie, J.: Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100**(1), 139–156 (2012). (11)
16. Cai, T., Liu, W., Zhou, H.H.: Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *Ann. Stat.* **44**(2), 455–488 (2016)
17. Cai, T.T., Ren, Z., Zhou, H.H.: Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electron. J. Stat.* **10**(1), 1–59 (2016)
18. Cai, T., Liu, W., Luo, X.: A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* **106**(494), 594–607 (2011)

19. Dahl, J., Vandenberghe, L., Roychowdhury, V.: Covariance selection for non-chordal graphs via chordal embedding. *Optim. Methods Softw.* **23**(4), 501–520 (2008)
20. d'Aspremont, A., Banerjee, O., El Ghaoui, L.: First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30**(1), 56–66 (2008)
21. Dempster, A.P.: Covariance selection. *Biometrics* **28**(1), 157–175 (1972)
22. El Karoui, N.: Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Stat.* **36**(6), 2717–2756 (2008)
23. Fan, J., Feng, Y., Yichao, W.: Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Stat.* **3**(2), 521–541 (2009)
24. Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441 (2007). (12)
25. Ghosh, S., Mukherjee, S. S.: Learning with latent group sparsity via heat flow dynamics on networks. [arXiv:2201.08326](https://arxiv.org/abs/2201.08326) (2022)
26. Huang, J.Z., Liu, N., Pourahmadi, M., Liu, L.: Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**(1), 85–98 (2006)
27. Huang, S., Li, J., Sun, L., Ye, J., Fleisher, A., Teresa, W., Chen, K., Reiman, E.: Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage* **50**(3), 935–949 (2010). <https://doi.org/10.1016/j.neuroimage.2009.12.120>
28. Kehler, J., Koehler, F., Meka, R., Moitra, A.: Learning some popular gaussian graphical models without condition number bounds. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 10986–10998. Curran Associates Inc, New York (2020)
29. Kehler, J., Koehler, F., Meka, R., Moitra, A.: Learning some popular gaussian graphical models without condition number bounds. *Adv. Neural. Inf. Process. Syst.* **33**, 10986–10998 (2020)
30. Lam, C., Fan, J.: Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.* **37**(6B), 4254–4278 (2009)
31. Lei, J., Rinaldo, A.: Consistency of spectral clustering in stochastic block models. *Ann. Stat.* **43**(1), 215–237 (2015). <https://doi.org/10.1214/14-AOS1274>
32. Liu, H., Lafferty, J., Wasserman, L.: The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**(80), 2295–2328 (2009)
33. Loh, P.-L., Bühlmann, P.: High-dimensional learning of linear causal networks via inverse covariance estimation. *J. Mach. Learn. Res.* **15**(1), 3065–3105 (2014)
34. Ma, Y., Garnett, R., Schneider, J.: σ -optimality for active learning on gaussian random fields. *Advances in Neural Information Processing Systems*, 26 (2013)
35. Malioutov, D.M., Johnson, J.K., Willsky, A.S.: Walk-sums and belief propagation in gaussian graphical models. *J. Mach. Learn. Res.* **7**(73), 2031–2064 (2006)
36. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* **34**(3), 1436–1462 (2006)
37. Menéndez, P., Kourmpetis, Y.A., ter Braak, C.J., van Eeuwijk, F.A.: Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the DREAM4 challenge. *PLoS ONE* **5**(12), e14147 (2010). <https://doi.org/10.1371/journal.pone.0014147>
38. Misra, S., Vuffray, M., Likhov, A. Y.: Information theoretic optimal learning of gaussian graphical models. In: Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2888–2909. PMLR, 09–12 (2020)
39. Müller, A., Scarsini, M.: Archimedean copulae and positive dependence. *J. Multivar. Anal.* **93**(2), 434–445 (2005). <https://doi.org/10.1016/j.jmva.2004.04.003>
40. Ravikumar, P., Wainwright, M.J., Lafferty, J.D.: High-dimensional ising model selection using ℓ_1 -regularized logistic regression (2010)
41. Ravikumar, P., Wainwright, M.J., Raskutti, G., Bin, Yu.: High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence. *Electron. J. Stat.* **5**(none), 935–980 (2011)
42. Rish, I., Thyreau, B., Thirion, B., Plaze, M., Paillere-martinot, M., Martelli, C., Martinot, J., Poline, J., Cecchi, G.: Discriminative network models of schizophrenia. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (eds.) *Advances in Neural Information Processing Systems*, vol. 22. Curran Associates Inc, New York (2009)
43. Rothman, A.J., Bickel, P.J., Levina, E., Zhu, J.: Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2**, 494–515 (2008)
44. Schafer, J., Strimmer, K.: Learning large-scale graphical Gaussian models from genomic data. In: *AIP Conference Proceedings*, pp. 263–276. AIP (2005). <https://doi.org/10.1063/1.1985393>
45. Sheffield, S.: Gaussian free fields for mathematicians. *Probab. Theory Relat. Fields* **139**(3–4), 521–541 (2007)

46. Shi, W., Ghosal, S., Martin, R.: Bayesian estimation of sparse precision matrices in the presence of Gaussian measurement error. *Electron. J. Stat.* **15**(2), 4545–4579 (2021)
47. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory* **52**(3), 1030–1051 (2006). <https://doi.org/10.1109/TIT.2005.864420>
48. Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., Thirion, B.: Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In: T. Jiang, N. Navab, J.P.W. Pluim, M.A. Viergever, (eds) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, pp. 200–208, Springer, Berlin (2010)
49. Varoquaux, G., Gramfort, A., Poline, J., Thirion, B.: Brain covariance selection: better individual functional connectivity models using population prior. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) *Advances in Neural Information Processing Systems*, vol. 23. Curran Associates Inc, New York (2010)
50. Vershynin, R.: *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47. Cambridge University Press, Cambridge (2018)
51. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory* **55**(5), 2183–2202 (2009). <https://doi.org/10.1109/TIT.2009.2016018>
52. Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelič, A., von Rohr, P., Thiele, L., Zitzler, E., Grusissem, W., Bühlmann, P.: Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.* **5**(11), R92 (2004). <https://doi.org/10.1186/gb-2004-5-11-r92>
53. Woodbury, M.A.: *Inverting Modified Matrices*. Princeton University, Department of Statistics, Princeton (1950)
54. Wu, W.B., Pourahmadi, M.: Non-parametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**(4), 831–844 (2003)
55. Yuan, M.: High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11**(79), 2261–2286 (2010)
56. Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35 (2007)
57. Zhao, P., Bin, Yu.: On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**(90), 2541–2563 (2006)
58. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919 (2003)
59. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In: *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, vol. 3 (2003)
60. Zwiernik, P.: *Semialgebraic Statistics and Latent Tree Models*. Monographs on Statistics and Applied Probability, vol. 146. Chapman & Hall/CRC, Boca Raton (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Subhro Ghosh¹ · Soumendu Sundar Mukherjee² · Hoang-Son Tran¹  · Ujan Gangopadhyay¹

Subhro Ghosh
subhrowork@gmail.com

Soumendu Sundar Mukherjee
ssmukherjee@isical.ac.in

Ujan Gangopadhyay
ujan@nus.edu.sg

- ¹ Department of Mathematics, National University of Singapore, Singapore, Singapore
- ² Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Calcutta, India