Generalized Score Matching for Non-Negative Data

Shiqing Yu sqyu@uw.edu
Mathias Drton md5@uw.edu
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Ali Shojaie ashojaie@uw.edu
Department of Biostatistics
University of Washington
Seattle, WA 98195-7232, USA

Editor:

Abstract

A common challenge in estimating parameters of probability density functions is the intractability of the normalizing constant. In such cases maximum likelihood estimation could in principle be implemented using numerical integration, but this approach becomes computationally prohibitive for larger scale problems. In contrast, the score matching method of Hyvärinen (2005) avoids direct calculation of the normalizing constant. The method is designed for continuous distributions supported on Euclidean space $\mathbb{R}^m$ and is particularly convenient for exponential families, for which it yields closed-form estimates. Hyvärinen (2007) extended the approach to distributions supported on the non-negative orthant $\mathbb{R}^m_+$. In this paper, we give a generalized form of score matching for non-negative data that improves estimation efficiency. We also generalize the regularized score matching method of Lin et al. (2016), giving a modification that avoids a general existence issue for the estimator and yields practically as well as theoretically improved estimators for non-negative Gaussian graphical models. This is an extended version of an earlier work by Yu et al. (2018).

Keywords: Exponential family, graphical model, positive data, score matching, sparsity.

1. Introduction

Graphical models specify conditional independence relations for a random vector $\boldsymbol{X} = (X_i)_{i\in V}$ indexed by the nodes of a graph (Lauritzen, 1996). For undirected graphs, variables X_i and X_j are required to be conditionally independent given $(X_k)_{k\neq i,j}$ if there is no edge between i and j. The smallest undirected graph with this property is the *conditional independence graph* of $\boldsymbol{X}$. Estimation of this graph and associated interaction parameters has been a topic of continued research as reviewed by Drton and Maathuis (2017).

Largely due to their tractability, Gaussian graphical models (GGMs) have gained great popularity. The conditional independence graph of a multivariate normal vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is determined by the *inverse covariance matrix* $\mathbf{K} \equiv \boldsymbol{\Sigma}^{-1}$, also termed *concentration* or *precision matrix*. Specifically, X_i and X_j are conditionally independent given all other variables if and only if the (i, j)-th and the (j, i)-th entry of $\mathbf{K}$ are both zero. This

simple relation underlies a rich literature including Drton and Perlman (2004), Meinshausen and Bühlmann (2006), Yuan and Lin (2007) and Friedman et al. (2008), among others.

More recent work has provided tractable procedures also for non-Gaussian graphical models. This includes Gaussian copula models (Liu et al., 2009; Dobra and Lenkoski, 2011; Liu et al., 2012), Ising models (Ravikumar et al., 2010), other exponential family models (Chen et al., 2014; Yang et al., 2015), as well as semi- or non-parametric estimation techniques (Fellinghauer et al., 2013; Voorman et al., 2013). In this paper, we focus on non-negative Gaussian random variables, as recently considered by Lin et al. (2016) and Yu et al. (2016). However, our main ideas can also be applied for other classes of distributions whose support is restricted to a rectangular set.

Let $\boldsymbol{X}$ be a random vector whose coordinates take non-negative values. Let $\boldsymbol{\mu} \in \mathbb{R}^p$, and let $\mathbf{K}$ be a positive definite matrix. Then $\boldsymbol{X}$ follows a truncated normal distribution with mean parameter $\boldsymbol{\mu}$ and inverse covariance parameter $\mathbf{K}$, in symbols $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{K})$, if it has (Lebesgue) density proportional to

$$\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{K}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1}$$

on $\mathbb{R}^m_+ \equiv [0,+\infty)^m$. We refer to $\boldsymbol{\Sigma} = \mathbf{K}^{-1}$ as the covariance parameter of the distribution. The conditional independence graph of truncated normal $\boldsymbol{X}$ is determined just as in the Gaussian case: X_i and X_j are conditionally independent given all other variables if and only if $\kappa_{ij} = \kappa_{ji} = 0$. In this paper, we develop estimators of $(\boldsymbol{\mu}, \mathbf{K})$ and the associated conditional independence graph based on the idea of score matching (Hyvärinen, 2005).

Score matching was first developed for continuous distributions supported on all of $\mathbb{R}^m$. Consider such a distribution P_0, with density p_0 and support $\mathbb{R}^m$, so $p_0(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \mathbb{R}^m$. Let $\mathcal{P}$ be a family of distributions with twice differentiable densities. The score matching estimator of p_0 using $\mathcal{P}$ as a model is the minimizer of the expected squared ℓ_2 distance between the gradients of $\log p_0$ and a log-density from $\mathcal{P}$. So we minimize the loss $\int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x})\|_2^2 \,\mathrm{d}\boldsymbol{x}$ with respect to densities p from $\mathcal{P}$. The loss depends on p_0 but partial integration can be used to rewrite it in a form that can be approximated by averaging over the sample without knowing p_0. A key feature of score matching is that normalizing constants cancel in gradients of log-densities, allowing for simple treatment of models with intractable normalizing constants. Furthermore, for exponential families, the loss is quadratic in the canonical parameter, making optimization straightforward.

If the considered distributions are supported on a proper subset of $\mathbb{R}^m$, then the partial integration arguments underlying the score matching estimator may fail due to discontinuities at the boundary of the support. For data supported on $\mathbb{R}^m_+$, Hyvärinen (2007) addresses this problem by modifying the loss to $\int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{x} - \nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{x}\|_2^2 \,\mathrm{d}\boldsymbol{x}$, where $\circ$ denotes entrywise multiplication. In this loss, boundary effects are dampened by multiplying gradients elementwise with the identity functions x_j. In this paper, we propose *generalized score matching* methods that are based on elementwise multiplication with functions other than x_j. As we show, this can lead to drastically improved estimation accuracy in both simulations and theory.

Lin et al. (2016) estimate truncated GGMs based on this modification, with an ℓ_1 penalty on the entries of $\mathbf{K}$ added to the loss. However, the paper overlooks the fact that the loss can be unbounded from below in the high-dimensional setting even with an ℓ_1 penalty, and

hence a solution may not exist. In fact, this issue also arises for the regularized version of the original score matching for Gaussian graphical models, which turns out to coincide with the estimators from Zhang and Zou (2014) and Liu and Luo (2015). Since the unpenalized loss is quadratic in the parameter to be estimated, we propose modifying it by adding small positive values to the diagonals of the positive semi-definite matrix that defines the quadratic part, in order to ensure the loss is bounded and strongly convex and thus admits a unique minimizer. We develop this idea for the original score matching estimator for GGMs and for the *generalized score matching* estimator for graphical models for non-negative data proposed in this paper. We show both theoretically and empirically that the consistency results still hold (or even improve) if the positive values added are smaller than a threshold that is readily computable.

The paper is organized as follows. Section 2 introduces score matching and our proposed *generalized score matching*. In Section 3, we apply generalized score matching to exponential families, with univariate truncated Gaussian distributions as an example. *Regularized generalized score matching* for graphical models is formulated in Section 4, where we also address the non-existence issue in Gaussian models. Section 5 gives a detailed analysis of truncated GGMs. Simulation results and applications to RNAseq data are given in Section 6. Section gives a brief discussion of our results. Proofs for theorems in Sections 3-5 are presented in Section 8 and Appendix A. Appendix B gives additional simulation results.

1.1 Notation

Constant scalars, vectors, and functions are written in lower-case (e.g., a, $\boldsymbol{a}$), random scalars and vectors in upper-case (e.g., X, $\boldsymbol{X}$). Regular font is used for scalars (e.g. a, X), and boldface for vectors (e.g. $\boldsymbol{a}$, $\boldsymbol{X}$). Matrices are in upright bold, with constant matrices in upper-case ($\mathbf{K}$, $\mathbf{M}$) and random matrices holding observations in lower-case ($\mathbf{x}$, $\mathbf{y}$). Subscripts refer to entries in vectors and columns in matrices. Superscripts refer to rows in matrices. So X_j is the j-th component of a random vector $\boldsymbol{X}$. For a data matrix $\mathbf{x} \in \mathbb{R}^{n\times m}$, each row comprising one observation of m variables/features, $X_j^{(i)}$ is the j-th feature for the i-th observation. Stacking the columns of a matrix $\mathbf{K} = [\kappa_{ij}]_{i,j} \in \mathbb{R}^{q\times r}$ gives its vectorization $\text{vec}(\mathbf{K}) = (\kappa_{11}, \ldots, \kappa_{q1}, \kappa_{12}, \ldots, \kappa_{q2}, \ldots, \kappa_{1r}, \ldots, \kappa_{qr})^\top$. For a matrix $\mathbf{K} \in \mathbb{R}^{q\times q}$, $\text{diag}(\mathbf{K}) \in \mathbb{R}^q$ denotes its diagonal, and for a vector $\boldsymbol{v} \in \mathbb{R}^q$, $\text{diag}(\boldsymbol{v})$ is the $q \times q$ diagonal matrix with diagonals $v_1, \ldots, v_q$.

For $a \geq 1$, the ℓ_a-norm of a vector $\boldsymbol{v} \in \mathbb{R}^q$ is denoted

$$\|\boldsymbol{v}\|_a = \left(\sum_{j=1}^{q} |v_j|^a \right)^{1/a},$$

with $\|\boldsymbol{v}\|_\infty = \max_{j=1,\ldots,q} |v_j|$. A matrix $\mathbf{K} = [\kappa_{ij}]_{i,j} \in \mathbb{R}^{q\times r}$ has Frobenius norm

$$|||\mathbf{K}|||_F \equiv \|\text{vec}(\mathbf{K})\|_2 \equiv \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{r} \kappa_{ij}^2},$$

and max norm $\|\mathbf{K}\|_\infty \equiv \|\mathrm{vec}(\mathbf{K})\|_\infty \equiv \max_{i,j} |\kappa_{ij}|$. Its ℓ_a-ℓ_b operator norm is

$$|||\mathbf{K}|||_{a,b} \equiv \max_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\mathbf{K}\boldsymbol{x}\|_b}{\|\boldsymbol{x}\|_a}$$

with shorthand notation $|||\mathbf{K}|||_a \equiv |||\mathbf{K}|||_{a,a}$; for instance, $|||\mathbf{K}|||_\infty \equiv \max_{i=1\ldots,q} \sum_{j=1}^{r} |\kappa_{ij}|$.

For a function $f : \mathbb{R}^m \to \mathbb{R}$, we define $\partial_j f(\boldsymbol{x})$ as the partial derivative with respect to x_j, and $\partial_{jj} f(\boldsymbol{x}) = \partial_j \partial_j f(\boldsymbol{x})$. For $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$, $\boldsymbol{f}(x) = (f_1(x), \ldots, f_m(x))^\top$, we let $\boldsymbol{f}'(x) = (f_1'(x), \ldots, f_m'(x))^\top$ be the vector of derivatives. Likewise $\boldsymbol{f}''(x)$ is used for second derivatives. The symbol $\mathbb{1}_A(\cdot)$ denotes the indicator function of the set A, while $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all 1's. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{a} \circ \boldsymbol{b} \equiv (a_1 b_1, \ldots, a_m b_m)^\top$. A density of a distribution is always a probability density function with respect to Lebesgue measure. When it is clear from the context, $\mathbb{E}_0$ denotes the expectation under a true distribution P_0.

2. Score Matching

2.1 Original Score Matching

Suppose $\boldsymbol{X}$ is a random vector taking values in $\mathbb{R}^m$ with distribution P_0 and density p_0. Suppose $P_0 \in \mathcal{P}$, a family of distributions with twice differentiable densities supported on $\mathbb{R}^m$. The *score matching loss* for $P \in \mathcal{P}$, with density p, is

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x}. \tag{2}$$

The gradients in (2) can be thought of as gradients with respect to a hypothetical location parameter, evaluated at the origin (Hyvärinen, 2005). The loss $J(P)$ is minimized if and only if $P = P_0$, which forms the basis for estimation of P_0. Importantly, since the loss depends on p only through its log-gradient, it suffices to know p up to a normalizing constant. Under mild conditions, (2) can be rewritten as

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[\partial_{jj} \log p(\boldsymbol{x}) + \frac{(\partial_j \log p(\boldsymbol{x}))^2}{2} \right] \mathrm{d}\boldsymbol{x}, \tag{3}$$

plus a constant independent of p. The integral in (3) can be approximated by a sample average, which avoids knowing the true density p_0 and provides a way to estimate p_0.

2.2 Generalized Score Matching for Non-Negative Data

When the true density p_0 is supported on a proper subset of $\mathbb{R}^m$, the integration by parts underlying the equivalence of (2) and (3) may fail due to discontinuity at the boundary. For distributions supported on the non-negative orthant $\mathbb{R}^m_+$, Hyvärinen (2007) addressed this issue by instead minimizing the *non-negative score matching loss*

$$J_+(P) = \int_{\mathbb{R}^m_+} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{x} - \nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{x}\|_2^2 \, \mathrm{d}\boldsymbol{x}. \tag{4}$$

This loss can be motivated via gradients with respect to a hypothetical scale parameter (Hyvärinen, 2007). Under mild conditions, $J_+(P)$ can again be rewritten in terms of an expectation of a function independent of p_0, thus allowing one to form a sample loss.

In this work, we consider generalizing the non-negative score matching loss as follows.

Definition 1 *Let $\mathcal{P}_+$ be the family of all distributions with twice differentiable densities supported on $\mathbb{R}^m_+$. Suppose the m-variate random vector $\boldsymbol{X}$ has true distribution $P_0 \in \mathcal{P}_+$, and let p_0 be its twice differentiable density. Let $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ be a.e. (almost everywhere) positive and a.e. differentiable functions, and set $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x_1), \ldots, h_m(x_m))^\top$. For $P \in \mathcal{P}_+$ with density p, the* generalized $\boldsymbol{h}$-score matching loss *is*

$$J_{\boldsymbol{h}}(P) = \int_{\mathbb{R}^m_+} \frac{1}{2} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} - \nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2}\|_2^2 \, \mathrm{d}\boldsymbol{x}, \tag{5}$$

where $\boldsymbol{h}^{1/2}(\boldsymbol{x}) \equiv (h_1^{1/2}(x_1), \ldots, h_m^{1/2}(x_m))^\top$.

Proposition 2 *The distribution P_0 is the unique minimizer of $J_{\boldsymbol{h}}(P)$ for $P \in \mathcal{P}_+$.*

Proof First, observe that $J_{\boldsymbol{h}}(P) \geq 0$ and $J_{\boldsymbol{h}}(P_0) = 0$. For uniqueness, suppose $J_{\boldsymbol{h}}(P_1) = 0$ for some $P_1 \in \mathcal{P}_+$. Let p_0 and p_1 be the respective densities. By assumption $p_0(\boldsymbol{x}) > 0$ a.s. (almost surely) and $h_j^{1/2}(\boldsymbol{x}) > 0$ a.s. for all $j = 1, \ldots, m$. Therefore, we must have $\nabla \log p_1(\boldsymbol{x}) = \nabla \log p_0(\boldsymbol{x})$ a.s., or equivalently, $p_1(\boldsymbol{x}) = \text{const} \times p_0(\boldsymbol{x})$ for almost every $\boldsymbol{x} \in \mathbb{R}^m_+$. Since p_1 and p_0 are both densities and continuous, it follows that $p_1(\boldsymbol{x}) = p_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^m_+$. ■

Choosing all $h_j(x) = x^2$ recovers the loss from (4). In our generalization, we will focus on using functions h_j that are increasing but are bounded or grow rather slowly. This will alleviate the need to estimate higher moments, leading to better practical performance and improved theoretical guarantees. We note that our approach could also be presented in terms of transformations of data; compare to Section 11 in Parry et al. (2012). In particular, log-transforming positive data into all of $\mathbb{R}^m$ and then applying (2) is equivalent to (4).

We will consider the following assumptions:

(A1) $\lim_{x_j \nearrow +\infty, x_j \searrow 0^+} p_0(\boldsymbol{x}) h_j(x_j) \partial_j \log p(\boldsymbol{x}) = 0, \quad \forall \boldsymbol{x}_{-j} \in \mathbb{R}^{m-1}_+, \quad \forall p \in \mathcal{P}_+,$

(A2) $\mathbb{E}_{p_0} \|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 < +\infty, \;\; \mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 < +\infty, \;\; \forall p \in \mathcal{P}_+,$

where $\forall p \in \mathcal{P}_+$ is a shorthand for "for all p being the density of some $P \in \mathcal{P}_+$", and the prime symbol denotes component-wise differentiation. While the second half of (A2) was not made explicit in Hyvärinen (2005) and Hyvärinen (2007), (A1)-(A2) were both required for integration by parts and Fubini-Tonelli to apply.

Once the forms of p_0 and p are given, sufficient conditions for $\boldsymbol{h}$ for Assumptions (A1)-(A2) to hold are easy to find. In particular, (A1) and (A2) are easily satisfied for exponential families: a sufficient condition for (A1) is bounded h_j and $\lim_{x \searrow 0^+} h_j(x) = 0$. This observation gives rise to the following requirements which will appear in our analysis.

Definition 3 *Suppose* $\boldsymbol{h} : \mathbb{R}_+^m \to \mathbb{R}_+^m$ *with* $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x_1), \ldots, h_m(x_m))^\top$. *We write that* $\boldsymbol{h} \in \mathcal{H}$ *(for simplicity we omit the dependency on* m*) if for all* $j = 1, \ldots, m$*:*

i) h_j *has derivative* h_j' *a.e. on* $\mathbb{R}_+$*;*

ii) $h_j(x) > 0$ *a.e. on* $\mathbb{R}_+$*;*

iii) $\lim_{x \searrow 0+} h_j(x) = 0$.

If in addition there are constants M *and* M' *such that* $0 \leq h_j(x) \leq M$ *and* $0 \leq h_j'(x) \leq M'$ *almost everywhere, we write* $\boldsymbol{h} \in \mathcal{H}_{M,M'}$.

Integration by parts yields the following theorem which shows that $J_{\boldsymbol{h}}$ from (5) is an expectation (under P_0) of a function that does not depend on p_0, similar to (3). The proof is given in Section 8.

Theorem 4 *Under (A1) and (A2), the loss from (5) equals*

$$J_{\boldsymbol{h}}(P) = \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^m \Big[h_j'(x_j) \partial_j (\log p(\boldsymbol{x})) + h_j(x_j) \partial_{jj} (\log p(\boldsymbol{x})) + \frac{1}{2} h_j(x_j) \left(\partial_j (\log p(\boldsymbol{x})) \right)^2 \Big] \, \mathrm{d}\boldsymbol{x} \quad (6)$$

plus a constant independent of p.

Given a data matrix $\mathbf{x} \in \mathbb{R}^{n \times m}$ with rows $\boldsymbol{X}^{(i)}$, we define the sample version of (6) as

$$\hat{J}_{\boldsymbol{h}}(P) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \left\{ h_j'(X_j^{(i)}) \partial_j (\log p(\boldsymbol{X}^{(i)})) + h_j(X_j^{(i)}) \left[\partial_{jj} (\log p(\boldsymbol{X}^{(i)})) + \frac{1}{2} \left(\partial_j (\log p(\boldsymbol{X}^{(i)})) \right)^2 \right] \right\}. \quad (7)$$

Subsequently, for a distribution P with density p, we let $J_{\boldsymbol{h}}(p) \equiv J_{\boldsymbol{h}}(P)$. Similarly, when a distribution $P_{\boldsymbol{\theta}}$ with density $p_{\boldsymbol{\theta}}$ is associated to a parameter vector $\boldsymbol{\theta}$, we write $J_{\boldsymbol{h}}(\boldsymbol{\theta}) \equiv J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) \equiv J_{\boldsymbol{h}}(P_{\boldsymbol{\theta}})$. We apply similar conventions to the sample version $\hat{J}_{\boldsymbol{h}}(P)$.

3. Exponential Families

In this section, we study the case where $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is an exponential family comprising continuous distributions with support $\mathbb{R}_+^m$. More specifically, we consider densities that are indexed by the canonical parameter $\boldsymbol{\theta} \in \mathbb{R}^r$ and have the form

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}_+^m, \quad (8)$$

where $\boldsymbol{t}(\boldsymbol{x}) \in \mathbb{R}_+^r$ comprises the sufficient statistics, $\psi(\boldsymbol{\theta})$ is a normalizing constant depending on $\boldsymbol{\theta}$ only, and $b(\boldsymbol{x})$ is the base measure, with $\boldsymbol{t}$ and b a.e. differentiable with respect to each component. Define $\boldsymbol{t_j'}(\boldsymbol{x}) \equiv (\partial_j t_1(\boldsymbol{x}), \ldots, \partial_j t_r(\boldsymbol{x}))^\top$ and $b_j'(\boldsymbol{x}) \equiv \partial_j b(\boldsymbol{x})$.

Theorem 5 *Under assumptions (A1)-(A2) from Section 2.2, the empirical generalized $\boldsymbol{h}$-score matching loss (7) can be rewritten as a quadratic function in $\boldsymbol{\theta} \in \mathbb{R}^r$:*

$$\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \text{const}, \quad \textit{where} \tag{9}$$

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top \quad \textit{and} \tag{10}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\left[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\right] \tag{11}$$

are sample averages of functions of the data matrix $\mathbf{x}$ only.

Define $\boldsymbol{\Gamma}_0 \equiv \mathbb{E}_{p_0}\boldsymbol{\Gamma}(\mathbf{x})$, $\boldsymbol{g}_0 \equiv \mathbb{E}_{p_0}\boldsymbol{g}(\mathbf{x})$, and $\boldsymbol{\Sigma}_0 \equiv \mathbb{E}_{p_0}[(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - g(\mathbf{x}))(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - g(\mathbf{x}))^\top]$.

Theorem 6 *Suppose that*

(C1) $\boldsymbol{\Gamma}$ *is invertible almost surely, and*

(C2) $\boldsymbol{\Gamma}_0$, $\boldsymbol{\Gamma}_0^{-1}$, $\boldsymbol{g}_0$ *and* $\boldsymbol{\Sigma}_0$ *exist and are finite entry-wise.*

Then the minimizer of (9) is a.s. unique with closed-form solution $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$, and

$$\hat{\boldsymbol{\theta}} \to_{a.s.} \boldsymbol{\theta}_0 \quad \textit{and} \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Gamma}_0^{-1}\right) \quad \textit{as } n \to \infty.$$

Theorem 5 clarifies the quadratic nature of the loss and is derived in Section 8. Theorem 6 gives asymptotic properties of the generalized $\boldsymbol{h}$-score matching estimator that in turn yield asymptotic tests and confidence intervals for $\boldsymbol{\theta}$. The theorem is an instance of Theorem 5.41 in van der Vaart (1998). Note that Condition (C1) holds if and only if $h_j(X_j) > 0$ a.e. and $[\boldsymbol{t}_j'(\boldsymbol{X}^{(1)}), \ldots, \boldsymbol{t}_j'(\boldsymbol{X}^{(n)})] \in \mathbb{R}^{r\times n}$ has rank r a.e. for some $j = 1, \ldots, m$.

In the following two examples, we assume (A1)-(A2) and (C1)-(C2).

3.1 Univariate Truncated Gaussian with Known Variance

Consider univariate ($m = r = 1$) truncated Gaussian distributions with unknown mean parameter μ and known variance parameter σ^2, so

$$p_\mu(x) \propto \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}_+. \tag{12}$$

In canonical form as in (8), we write

$$p_\theta(x) \propto \exp\left\{\theta t(x) + b(x)\right\}, \quad \theta \equiv \frac{\mu}{\sigma^2}, \quad t(x) \equiv x, \quad b(x) = -\frac{x^2}{2\sigma^2}.$$

Theorem 7 *Given i.i.d. samples $X_1, \ldots, X_n \sim p_{\mu_0}$, the generalized h-score matching estimator of μ is*

$$\hat{\mu}_h \equiv \frac{\sum_{i=1}^n h(X_i)X_i - \sigma^2 h'(X_i)}{\sum_{i=1}^n h(X_i)}.$$

If $\lim_{x\searrow 0^+} h(x) = 0$, $\lim_{x\nearrow+\infty} h^2(x)(x-\mu_0)p_{\mu_0}(x) = 0$ *and the expectations are finite (for example, when* $h(x) = o(\exp(Mx^2))$ *for* $M < \frac{1}{4\sigma^2}$*), then*

$$\sqrt{n}(\hat{\mu}_h - \mu_0) \to_d \mathcal{N}\left(0, \frac{\mathbb{E}_0[\sigma^2 h^2(X) + \sigma^4 h'^2(X)]}{\mathbb{E}_0^2[h(X)]}\right).$$

Moreover, the Cramér-Rao lower bound *for estimating* μ *is*

$$\frac{\sigma^4}{\mathrm{var}(X-\mu_0)}.$$

3.2 Univariate Truncated Gaussian with Known Mean

Consider univariate truncated Gaussian distributions with known mean parameter μ and unknown variance parameter $\sigma^2 > 0$, so

$$p_{\sigma^2}(x) \propto \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}_+.$$

In canonical form as in (8), we write

$$p_\theta(x) \propto \exp\left\{\theta t(x) + b(x)\right\}, \quad \theta \equiv \frac{1}{\sigma^2}, \quad t(x) \equiv -(x-\mu)^2/2, \quad b(x) = 0.$$

Theorem 8 *Given i.i.d. samples* $X_1, \ldots, X_n \sim p_{\sigma_0^2}$*, the generalized h-score matching estimator of* σ^2 *is*

$$\hat{\sigma}_h^2 \equiv \frac{\sum_{i=1}^n h(X_i)(X_i-\mu)^2}{\sum_{i=1}^n h(X_i) + h'(X_i)(X_i-\mu)}.$$

If, in addition to the assumptions in Theorem 7, $\lim\limits_{x\nearrow+\infty} h^2(x)(x-\mu)^3 p_{\sigma_0^2}(x) = 0$*, then*

$$\sqrt{n}(\hat{\sigma}_h^2 - \sigma_0^2) \to_d \mathcal{N}\left(0, \frac{2\sigma_0^6\mathbb{E}_0[h^2(X)(X-\mu)^2] + \sigma_0^8\mathbb{E}_0[h'^2(X)(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right).$$

Moreover, the Cramér-Rao lower bound *for estimating* σ^2 *is*

$$\frac{4\sigma_0^8}{\mathrm{var}(X-\mu)^2}.$$

Remark 9 If $\mu_0 = 0$, then $h(x) \equiv 1$ also satisfies (A1)-(A2) and (C1)-(C2) and one recovers the sample variance $\frac{1}{n}\sum_i X_i^2$, which obtains the Cramér-Rao lower bound.

The benefits of using bounded h in the above example can be explained as follows. When $\mu \gg \sigma$, there is effectively no truncation to the Gaussian distribution, and our method adapts to using low moments in (5), since a bounded and increasing $h(x)$ becomes almost constant as it reaches its asymptote for x large. Hence, we effectively revert to the original score matching (recall Section 2.1). In the other cases, the truncation effect is significant and our estimator uses higher moments accordingly.

3.3 Simulations for Univariate Truncated Normal Distributions

Figure 1 plots the asymptotic variance of $\hat{\mu}_h$ from Theorem 7, with $\sigma = 1$ known. Efficiency as measured by the Cramér-Rao lower bound divided by the asymptotic variance is also shown. We see that two truncated versions of $\log(1 + x)$ have asymptotic variance close to the Cramér-Rao bound. This asymptotic variance is also reflective of the variance for smaller finite samples.

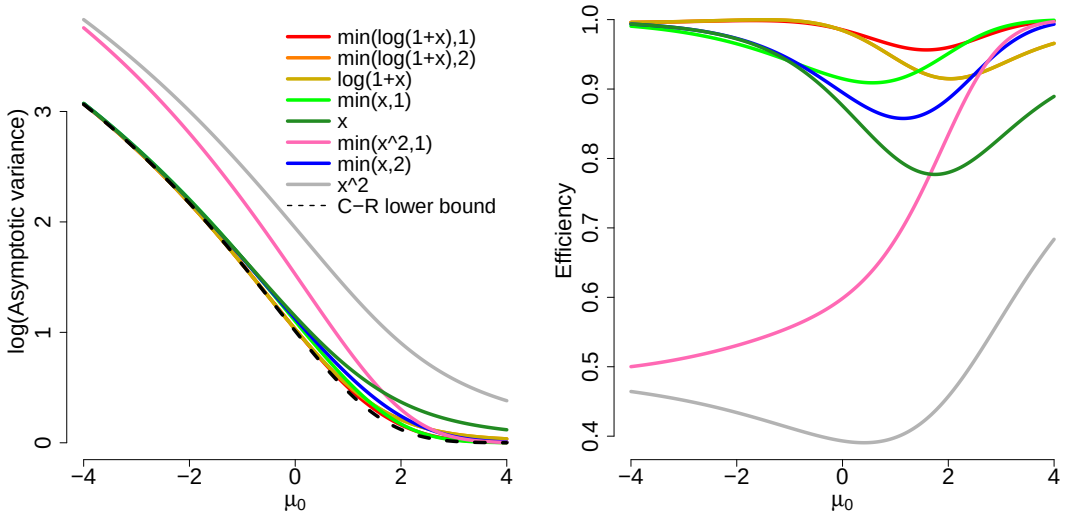


Figure 1: Log of asymptotic variance and efficiency with respect to the Cramér-Rao bound for $\hat{\mu}_h$ ($\sigma^2 = 1$ known).

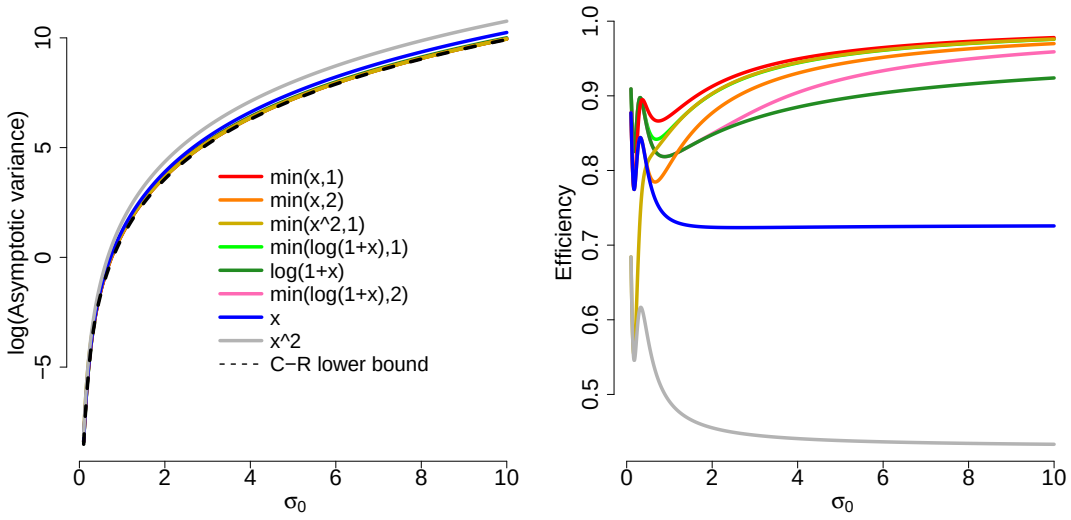


Figure 2: Log of asymptotic variance and efficiency with respect to the Cramér-Rao bound for $\hat{\sigma}_h^2$ ($\mu = 0.5$ known).

Figure 2 is the analog of Figure 1 for $\hat{\sigma}_h^2$ with $\mu = 0.5$ known. While the specifics are a bit different the benefits of using bounded or slowly growing h are again clear. We note that when σ is small, the effect of truncation to the positive part of the real line is small.

4. Regularized Generalized Score Matching

It is well-known that in high-dimensional settings, where the number r of parameters to estimate may be larger than the sample size n, it is hard, if not impossible, to estimate the parameters consistently, without turning to some form of regularization. More specifically, for exponential families, condition (C1) in Section 3 fails when $r > n$. A popular approach is then the use of ℓ_1 regularization to exploit possible sparsity.

4.1 Ensuring Bounded ℓ_1-Regularized Loss

Let the data matrix $\mathbf{x} \in \mathbb{R}^{n\times m}$ comprise n i.i.d. samples from distribution P_0. Assume P_0 has density p_0 belonging to an exponential family $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \subset \mathcal{P}_+$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^r$. Adding an ℓ_1 penalty to (9), we obtain the regularized generalized score matching loss

$$\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1 \tag{13}$$

as in Lin et al. (2016). The loss in (13) involves a quadratic smooth part as in the familiar lasso loss for linear regression. However, although the matrix $\boldsymbol{\Gamma}$ is positive semidefinite, the regularized loss in (13) is not guaranteed to be bounded unless the tuning parameter λ is sufficiently large—a problem that does not occur in lasso. We note that here, and throughout, we suppress the dependence on the data $\mathbf{x}$ for $\boldsymbol{\Gamma}(\mathbf{x})$, $\boldsymbol{g}(\mathbf{x})$ and derived quantities.

For a more detailed explanation, note that that by (10), $\boldsymbol{\Gamma} = \mathbf{H}^\top\mathbf{H}$ for some $\mathbf{H} \in \mathbb{R}^{nm\times r}$. In the high-dimensional case, the rank of $\boldsymbol{\Gamma}$, or equivalently $\mathbf{H}$, is at most $nm < r$. Hence, $\boldsymbol{\Gamma}$ is not invertible and $\boldsymbol{g}$ does not necessarily lie in the column span of $\boldsymbol{\Gamma}$. Let $\mathrm{Ker}(\boldsymbol{\Gamma})$ be the kernel of $\boldsymbol{\Gamma}$. Then there may exist $\boldsymbol{\nu} \in \mathrm{Ker}(\boldsymbol{\Gamma})$ with $\boldsymbol{g}^\top\boldsymbol{\nu} \neq 0$. In this case, if

$$0 < \lambda < \sup_{\boldsymbol{\nu}\in\mathrm{Ker}(\boldsymbol{\Gamma})} |\boldsymbol{g}^\top\boldsymbol{\nu}|/\|\boldsymbol{\nu}\|_1,$$

there exists $\boldsymbol{\nu} \in \mathrm{Ker}(\boldsymbol{\Gamma})$ with $\frac{1}{2}\boldsymbol{\nu}^\top\boldsymbol{\Gamma}\boldsymbol{\nu} = 0$ and $-\boldsymbol{g}^\top\boldsymbol{\nu} + \lambda\|\boldsymbol{\nu}\|_1 < 0$. Evaluating at $\boldsymbol{\theta}(a) = a\cdot\boldsymbol{\nu}$ for scalar $a > 0$, the loss becomes $a\left(-\boldsymbol{g}^\top\boldsymbol{\nu} + \lambda\|\boldsymbol{\nu}\|_1\right)$, which is negative and linear in a, and thus unbounded below. In this case no minimizer of (13) exists for small values of λ. This issue also exists for the estimators from Zhang and Zou (2014) and Liu and Luo (2015), which correspond to score matching for GGMs. We note that in the context of (truncated) GGMs the condition $nm < r$ reduces to $n < m$ as $r = m^2$.

To circumvent this problem we add small values $\gamma_\ell > 0$ to the diagonal entries of $\boldsymbol{\Gamma}$, which become $\boldsymbol{\Gamma}_{\ell,\ell} + \gamma_\ell$, $\ell = 1, \ldots, r$. This is in the spirit of work such as Ledoit and Wolf (2004) and corresponds to an elastic net-type penalty (Zou and Hastie, 2005) with weighted ℓ_2 penalty $\sum_{\ell=1}^r \gamma_\ell\theta_\ell^2$. After this modification $\boldsymbol{\Gamma}$ is positive definite, our regularized loss is strongly convex in $\boldsymbol{\theta}$, and a unique minimizer exists for all $\lambda \geq 0$. When applied to the special case of truncated GGMs, we show that a result on consistent estimation holds if we choose $\gamma_\ell = c\boldsymbol{\Gamma}_{\ell,\ell}$ for a suitably small constant $c > 0$, for which we propose a particular choice to avoid tuning. This choice of γ_ℓ depends on the data through $\boldsymbol{\Gamma}_{\ell,\ell}$.

Definition 10 *For* $\boldsymbol{\gamma} \in \mathbb{R}_+^r \backslash \{\mathbf{0}\}$, *let* $\boldsymbol{\Gamma}_{\boldsymbol{\gamma}} \equiv \boldsymbol{\Gamma} + \mathrm{diag}(\boldsymbol{\gamma})$. *The* regularized generalized $\boldsymbol{h}$-score matching estimator *with tuning parameter* $\lambda \geq 0$ *and* amplifiers $\boldsymbol{\gamma}$ *is the estimator*

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma}}(\boldsymbol{\theta}) \equiv \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) \boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1. \tag{14}$$

We note that $\hat{\boldsymbol{\theta}}$ from (14) is a *piecewise linear* function of λ (Lin et al., 2016).

4.2 Theory for Graphical Models

The graphical models we treat are parametrized by an inverse covariance matrix and possibly a mean vector. It is convenient to accommodate this setting with a matrix-valued parameter $\boldsymbol{\Psi} \in \mathbb{R}^{r_1 \times r_2}$ (in place of $\boldsymbol{\theta}$) and specifying our regularized $\boldsymbol{h}$-score matching loss as

$$\hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma}}(\boldsymbol{\Psi}) \equiv \underset{\boldsymbol{\Psi} \in \mathbb{R}^{r_1 \times r_2}}{\operatorname{argmin}} \frac{1}{2} \mathrm{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) \mathrm{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}(\mathbf{x})^\top \mathrm{vec}(\boldsymbol{\Psi}) + \lambda \|\boldsymbol{\Psi}\|_1. \tag{15}$$

For truncated centered GGMs, $\boldsymbol{\Psi}$ is the $m \times m$ inverse covariance matrix $\mathbf{K}$. For truncated non-centered GGMs we take $\boldsymbol{\Psi} = [\mathbf{K}, \mathbf{K}\boldsymbol{\mu}]^\top$, where $\boldsymbol{\mu}$ is the mean parameter. Following related prior work such as Lin et al. (2016), we allow the matrix $\mathbf{K}$ to be nonsymmetric, which allows us to decouple optimization over the different columns of $\mathbf{K}$ or $\boldsymbol{\Psi}$.

Definition 11 *Let* $\boldsymbol{\Gamma}_0 \equiv \mathbb{E}_0 \boldsymbol{\Gamma}(\mathbf{x})$ *and* $\boldsymbol{g}_0 \equiv \mathbb{E}_0 \boldsymbol{g}(\mathbf{x})$ *be the population versions of* $\boldsymbol{\Gamma}(\mathbf{x})$ *and* $\boldsymbol{g}(\mathbf{x})$ *under the distribution given by a true parameter matrix* $\boldsymbol{\Psi}_0$. *The support of a matrix* $\boldsymbol{\Psi}$ *is* $S(\boldsymbol{\Psi}) \equiv \{(i,j) : \psi_{ij} \neq 0\}$, *and we let* $S_0 = S(\boldsymbol{\Psi}_0)$. *For a matrix* $\boldsymbol{\Psi}_0$, *we define* $d_{\boldsymbol{\Psi}_0}$ *to be the maximum number of non-zero entries in any column, and* $c_{\boldsymbol{\Psi}_0} \equiv |||\boldsymbol{\Psi}_0|||_{\infty,\infty}$. *Writing* $\boldsymbol{\Gamma}_{0,AB}$ *for the* $A \times B$ *submatrix of* $\boldsymbol{\Gamma}_0$, *we define*

$$c_{\boldsymbol{\Gamma}_0} \equiv |||(\boldsymbol{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty}.$$

Finally, $\boldsymbol{\Gamma}_0$ *satisfies* the irrepresentability condition with incoherence parameter $\alpha \in (0,1]$ *if*

$$|||\boldsymbol{\Gamma}_{0,S_0^c S_0}(\boldsymbol{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty} \leq (1-\alpha). \tag{16}$$

Our analysis of the regularized generalized $\boldsymbol{h}$-score matching estimator builds on the following theorem taken from Lin et al. (2016, Theorem 1).

Theorem 12 *Suppose* $\boldsymbol{\Gamma}_0$ *has* $\boldsymbol{\Gamma}_{0,S_0S_0}$ *invertible and satisfies the irrepresentability condition (16) with incoherence parameter* $\alpha \in (0,1]$. *Assume that*

$$\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty < \epsilon_1, \qquad \|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2, \tag{17}$$

with $d_{\boldsymbol{\Psi}_0}\epsilon_1 \leq \alpha/(6c_{\boldsymbol{\Gamma}_0})$. *If*

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{\boldsymbol{\Psi}_0}\epsilon_1, \epsilon_2\},$$

then the following holds:

(a) The regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\boldsymbol{\Psi}}$ minimizing (15) is unique, with support $\hat{S} \equiv S(\hat{\boldsymbol{\Psi}}) \subseteq S_0$, and satisfies

$$\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda.$$

(b) If

$$\min_{1\leq j<k\leq m} |\boldsymbol{\Psi}_{0,jk}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda,$$

then $\hat{S} = S_0$ and $\operatorname{sign}(\hat{\boldsymbol{\Psi}}_{jk}) = \operatorname{sign}(\boldsymbol{\Psi}_{0.jk})$ for all $(j,k) \in S_0$.

This result is deterministic, and the improvement of our generalized estimator over the one in Lin et al. (2016) is in its probabilistic guarantees, as shown for truncated GGMs in Theorems 15 and 16 below. Before going into these examples, we state a general corollary.

Corollary 13 *Under the assumptions of Theorem 12, the matrix $\hat{\boldsymbol{\Psi}}$ minimizing (15) satisfies*

$$\begin{aligned}\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\|_F &\leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|} \leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda\sqrt{d_{\boldsymbol{\Psi}_0}m},\\ \|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\|_2 &\leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda\min(\sqrt{|S_0|}, d_{\boldsymbol{\Psi}_0}).\end{aligned}$$

4.3 Revisiting Gaussian Score Matching

In this section we briefly digress from score matching for non-negative data, and instead consider Example 1 from Lin et al. (2016) which applies the original score matching of Hyvärinen (2005) to centered Gaussian distributions. The density is of course proportional to $\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\mathbf{K}\boldsymbol{x}\right)$ for $\boldsymbol{x} \in \mathbb{R}^m$, and adding an ℓ_1 penalty gives the loss

$$\frac{1}{2}\operatorname{tr}\left(\mathbf{K}\mathbf{K}\cdot\frac{1}{n}\mathbf{x}\mathbf{x}^\top\right) - \operatorname{tr}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \tag{18}$$

which can be written as in (13) with $\boldsymbol{\theta} = \operatorname{vec}(\mathbf{K})$,

$$\boldsymbol{\Gamma} = \frac{1}{n}\operatorname{diag}(\mathbf{x}\mathbf{x}^\top, \ldots, \mathbf{x}\mathbf{x}^\top) \quad \text{and} \quad \boldsymbol{g} = \operatorname{vec}(\mathbf{I}_m).$$

The kernel of $\boldsymbol{\Gamma}$ need not be orthogonal to $\boldsymbol{g}$, and for λ small the loss can be unbounded below as discussed above. Corollary 1 of Lin et al. (2016) proves that the consistency results in Theorem 12 hold with probability $1 - m^{2-\tau}$ if $n > \max(c^* c_1^2 d_{\mathbf{K}}^2, 2)(\tau\log m + \log 4)$ and $\lambda > O(\sqrt{c^*(\tau\log m + \log 4)/n})$ for some constants c^* and c_1. However, this result is based on a tacit assumption of existence of the estimator. Our new modification resolves the issue as shown in the following theorem whose proof is given in Appendix A.

Theorem 14 *Adopt the amplifying strategy from Section 4.1 and redefine (18) as*

$$\frac{1}{2}\operatorname{tr}(\mathbf{K}\mathbf{K}\mathbf{G}) - \operatorname{tr}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \quad \mathbf{G}_{jk} = \frac{1}{n}(\mathbf{x}\mathbf{x}^\top)_{jk}(\mathbb{1}_{j\neq k} + c\cdot\mathbb{1}_{j=k}),$$

where $1 < c < 2 - \left(1 + 80\sqrt{\log m/n}\right)^{-1}$. Then the consistency result asserted in Corollary 1 of Lin et al. (2016) holds when replacing the constant c^* by $4c^*$.

5. Graphical Models for Truncated Gaussian Observations

5.1 Truncated Centered GGMs

Suppose now that $\boldsymbol{X} \sim \mathrm{TN}(\mathbf{0}, \mathbf{K})$ follows a truncated centered normal distribution. Our focus is then on estimating the inverse covariance/precision matrix parameter $\mathbf{K}$, which determines the conditional independence graph through its support $S(\mathbf{K}) \equiv \{(i,j) : \kappa_{ij} \neq 0\}$. This setting matches that of (15) with $\boldsymbol{\Psi} = \mathbf{K} \in \mathbb{R}^{m \times m}$.

For generalized score matching, consider a function $\boldsymbol{h} \in \mathcal{H}_{M,M'}$ as specified in Definition 3. In particular, the coordinate functions h_j and their derivatives h_j' are bounded a.e. by M and M', respectively. Boundedness is assumed for ease of proof in the main theorems. In our numerical experiments we will also consider slowly growing unbounded functions $\boldsymbol{h} \in \mathcal{H}$. Then, (A1)-(A2) are satisfied and the loss can be written as in (15). The j^{th} block of the $m^2 \times m^2$ block-diagonal non-amplified matrix $\boldsymbol{\Gamma}(\mathbf{x})$ is

$$\boldsymbol{\Gamma}_j(\mathbf{x}) \;\equiv\; \frac{1}{n}\mathbf{x}^\top \mathrm{diag}(\boldsymbol{h}_j(\boldsymbol{X}_j))\mathbf{x}, \tag{19}$$

where $\boldsymbol{h}_j(\boldsymbol{X}_j) \equiv [h_j(X_j^{(1)}), \ldots, h_j(X_j^{(n)})]^\top$; recall (10) and the form of truncated normal density in (1). Moreover, with $\mathbf{h}(\mathbf{x}) \equiv [h_j(X_j^{(i)})]_{i,j}$ and $\mathbf{h}'(\mathbf{x}) \equiv [h_j'(X_j^{(i)})]_{i,j}$,

$$\boldsymbol{g}(\mathbf{x}) \equiv \mathrm{vec}\left(\tfrac{1}{n}\mathbf{h}'(\mathbf{x})^\top \mathbf{x} + \mathrm{diag}\left(\tfrac{1}{n}\mathbf{h}(\mathbf{x})^\top \mathbf{1}_n\right)\right). \tag{20}$$

From Theorem 12, we may now derive the following probabilistic result about the regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\mathbf{K}} = \hat{\boldsymbol{\Psi}}$.

Theorem 15 *Suppose the data matrix* $\mathbf{x}$ *holds* n *i.i.d. copies of* $\boldsymbol{X} \sim \mathrm{TN}(\mathbf{0}, \mathbf{K}_0)$. *Assume that* $\boldsymbol{h} \in \mathcal{H}_{M,M'}$ *for constants* M, M', *and choose* $\boldsymbol{\gamma} = (c-1)\mathrm{diag}(\boldsymbol{\Gamma})$ *with*

$$1 < c < C(n,m) \equiv 2 - \left(1 + 4e\max\{6\log m/n, \sqrt{6\log m/n}\}\right)^{-1}.$$

Suppose further that $\boldsymbol{\Gamma}_0$ *has* $\boldsymbol{\Gamma}_{0,S_0S_0}$ *invertible and satisfies the irrepresentability condition (16) with* $\alpha \in (0,1]$. *Define* $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\mathbb{E}_0 X_j)$. *If for* $\tau > 3$ *the sample size and the regularization parameter satisfy*

$$n \geq \mathcal{O}\left(d_{\mathbf{K}_0}^2 \tau \log m \max\left\{\frac{c_{\boldsymbol{\Gamma}_0}^2 c_{\boldsymbol{X}}^4}{\alpha^2}, 1\right\}\right), \tag{21}$$

$$\lambda > \mathcal{O}\left[(c_{\mathbf{K}_0} c_{\boldsymbol{X}}^2 + c_{\boldsymbol{X}} + 1)\left(\sqrt{\frac{\tau\log m}{n}} + \frac{\tau\log m}{n}\right)\right], \tag{22}$$

then the following statements hold with probability $1 - m^{3-\tau}$:

(a) The regularized generalized $\boldsymbol{h}$*-score matching estimator* $\hat{\mathbf{K}}$ *that minimizes (15) is unique, has its support included in the true support,* $\hat{S} \equiv S(\hat{\mathbf{K}}) \subseteq S_0$, *and*

$$\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty < \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda,$$

$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|},$$
$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda \min(\sqrt{|S_0|}, d_{\mathbf{K}_0}).$$

(b) Moreover, if

$$\min_{j,k:(j,k)\in S_0} |\kappa_{0,jk}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda,$$

then $\hat{S} = S_0$ *and* $\operatorname{sign}(\hat{\kappa}_{jk}) = \operatorname{sign}(\kappa_{0,jk})$ *for all* $(j,k) \in S_0$.

The theorem is proved in Section 8, where details on the dependencies on constants are provided. A key ingredient of the proof is a tail bound on $\|\mathbf{\Gamma}_{\boldsymbol{\gamma}} - \mathbf{\Gamma}_0\|_\infty$, which features products of the $X_j^{(i)}$'s. In Lin et al. (2016), the products are up to fourth order. Using bounded $\boldsymbol{h}$ our products automatically calibrate to a quadratic polynomial when the observed values are large, and resort to higher moments only when they are small. This leads to improved bounds and convergence rates, underscored in the new requirement on the sample size n, which should be compared to $n \geq \mathcal{O}(d_{\mathbf{K}_0}^2(\log m^\tau)^8)$ in Lin et al. (2016).

We emphasize that it is indeed necessary to introduce amplifiers $\boldsymbol{\gamma} \succ \mathbf{0}$ or a multiplier $c > 1$ in addition to the ℓ_1 penalty. It is clear from (19) that $\operatorname{rank}(\mathbf{\Gamma}_j) \leq \operatorname{rank}(\mathbf{x}) \leq \min\{n, m\}$. Thus, $\mathbf{\Gamma}$ is non-invertible when $n < m$ and $\boldsymbol{g}$ need not lie in its column span. Note also that the vector $\boldsymbol{h}_j(\boldsymbol{X}_j)$ in (19) a.s. has all entries positive; recall that $h_j(x) > 0$ for $x > 0$ and $X_j > 0$ with probability 1. Hence, a.s., $\operatorname{rank}(\mathbf{\Gamma}_j) = \operatorname{rank}(\mathbf{x}) = \min\{n, m\}$.

5.2 Truncated Non-centered GGMs

Suppose now that $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}_0, \mathbf{K}_0)$ with both $\boldsymbol{\mu}_0$ and $\mathbf{K}_0$ unknown. The considered truncated normal model then comprises densities

$$p_{\boldsymbol{\mu},\mathbf{K}}(\boldsymbol{x}) \;\propto\; \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{K}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \mathbb{1}_{[0,\infty)^m}(\boldsymbol{x}). \tag{23}$$

Our focus is still on the inverse covariance parameter $\mathbf{K}$. With $\boldsymbol{\mu}$ also unknown, it will be convenient to work with the canonical parameters $\mathbf{K}$ and $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$. Estimates of $\boldsymbol{\mu}$ can be recovered using that $\boldsymbol{\mu} = \mathbf{K}^{-1}\boldsymbol{\eta}$.

Let $\mathbf{\Psi} \equiv [\mathbf{K}, \boldsymbol{\eta}]^\top \in \mathbb{R}^{(m+1)\times m}$ with columns $\mathbf{\Psi}_j = [\boldsymbol{k}_j, \eta_j]^\top$. Suppose again that $\boldsymbol{h} \in \mathcal{H}_{M,M'}$; recall Definition 3. Then (A1)-(A2) are satisfied and the unpenalized score-matching loss can be written as

$$\frac{1}{2}\operatorname{vec}(\mathbf{\Psi})^\top \mathbf{\Gamma}(\mathbf{x})\operatorname{vec}(\mathbf{\Psi}) - \boldsymbol{g}(\mathbf{x})^\top \operatorname{vec}(\mathbf{\Psi}).$$

The matrix $\mathbf{\Gamma}(\mathbf{x}) \in \mathbb{R}^{m(m+1)\times m(m+1)}$ is block-diagonal with j-th $(m+1)\times(m+1)$ block being

$$\mathbf{\Gamma}_j(\mathbf{x}) \equiv \begin{bmatrix} \mathbf{\Gamma}_{11,j} & \mathbf{\Gamma}_{12,j} \\ \mathbf{\Gamma}_{12,j}^\top & \mathbf{\Gamma}_{22,j} \end{bmatrix}, \tag{24}$$

where $\mathbf{\Gamma}_{11,j}$ is the $m \times m$ matrix defined in (19), $\mathbf{\Gamma}_{12,j} = -\frac{1}{n}\mathbf{x}^\top \boldsymbol{h}_j(\boldsymbol{X}_j)$, and $\mathbf{\Gamma}_{22,j} = \frac{1}{n}\sum_{i=1}^n h_j(X_j^{(i)})$. Partition the vector $\boldsymbol{g}(\mathbf{x})$ into m subvectors $\boldsymbol{g}_j(\mathbf{x}) \in \mathbb{R}^{m+1}$, where the

entries of $\boldsymbol{g}_j(\mathbf{x})$ correspond to column $\boldsymbol{\Psi}_j$. Then the k-th entry $\boldsymbol{g}_j(\mathbf{x})$ is

$$\boldsymbol{g}_{jk}(\mathbf{x}) \equiv \begin{cases} \frac{1}{n}\boldsymbol{h}_j(\boldsymbol{X}_j)\boldsymbol{X}_k & \text{if } k \le m,\ k \ne j, \\ \frac{1}{n}\boldsymbol{h}_j(\boldsymbol{X}_j)\boldsymbol{X}_k + \frac{1}{n}\sum_{i=1}^n h_j(X_j^{(i)}) & \text{if } k = j, \\ -\frac{1}{n}\sum_{i=1}^n h_j'(X_j^{(i)}) & \text{if } k = m+1. \end{cases} \tag{25}$$

Note that the entries for $k \le m$ also appear in (20).

By definition, $c_{\boldsymbol{\Psi}_0} \equiv |||\boldsymbol{\Psi_0}^\top|||_{\infty,\infty} \le c_{\mathbf{K}_0} + \|\boldsymbol{\eta}_0\|_\infty$, $d_{\boldsymbol{\Psi}_0} \le d_{\mathbf{K}_0} + 1$. The proof given for Theorem 15 goes through again here, and we have the following consistency results.

Theorem 16 *Suppose the data matrix holds n i.i.d. copies of $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}_0, \mathbf{K}_0)$. Assume that $\boldsymbol{h} \in \mathcal{H}_{M,M'}$ for constants M, M'. Let $\boldsymbol{\gamma}$ be a vector of amplifiers that are non-zero only for the diagonal entries of the matrices $\boldsymbol{\Gamma}_{11,j}$, amplifying those by $(c-1)\mathrm{diag}(\boldsymbol{\Gamma}_{11,j})$ for*

$$1 < c < C(n,m) \equiv 2 - \left(1 + 4e\max\{6\log m/n, \sqrt{6\log m/n}\}\right)^{-1}.$$

Suppose further that $\boldsymbol{\Gamma}_{0,S_0S_0}$ is invertible and satisfies the irrepresentability condition (16) with $\alpha \in (0,1]$. Define $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\mathbb{E}_0 X_j)$. If for $\tau > 3$ the sample size and the regularization parameter satisfy

$$n \ge \mathcal{O}\left(d_{\boldsymbol{\Psi}_0}^2 \tau \log m \max\left\{\frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}^2 c_{\boldsymbol{X}}^4}{\alpha^2}, 1\right\}\right), \tag{26}$$

$$\lambda > \mathcal{O}\left[(c_{\boldsymbol{\Psi}_0}c_{\boldsymbol{X}}^2 + c_{\boldsymbol{X}} + 1)\left(\sqrt{\frac{\tau\log m}{n}} + \frac{\tau\log m}{n}\right)\right], \tag{27}$$

then the following statements hold with probability $1 - m^{3-\tau}$:

(a) The regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\boldsymbol{\Psi}}$ that minimizes (15) is unique, has its support included in the true support, $\hat{S} \equiv S(\hat{\boldsymbol{\Psi}}) \subseteq S_0$, and satisfies

$$\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty < \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda, \qquad \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|_\infty < \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda,$$

$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F \le \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda\sqrt{|S_0|}, \qquad |||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|||_F \le \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda\sqrt{|S_0|},$$

$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 \le \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda\min(\sqrt{|S_0|}, d_{\boldsymbol{\Psi}_0}), \qquad |||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|||_2 \le \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda\min(\sqrt{|S_0|}, d_{\boldsymbol{\Psi}_0}).$$

(b) Moreover, if

$$\min_{j,k:(j,k)\in S_0}|\kappa_{0,jk}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda \qquad \textit{and} \qquad \min_{j:(m+1,j)\in S_0}|\eta_{0,j}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda,$$

then $\hat{S} = S_0$ and $\mathrm{sign}(\hat{\kappa}_{jk}) = \mathrm{sign}(\kappa_{0,jk})$ for all $(j,k) \in S_0$ and $\mathrm{sign}(\hat{\eta}_j) = \mathrm{sign}(\eta_{0j})$ for $(m+1,j) \in S_0$.

It is clear that amplification is needed to guarantee existence of a solution, similarly to what we observed in the centered case from Section 5.1. In Theorem 16, we considered positive amplifiers only for the submatrices $\Gamma_{11,j}$, which we claim is sufficient for unique existence of a solution for all penalty parameters $\lambda \geq 0$. To see this, consider any nonzero vector $\boldsymbol{\nu} \in \mathbb{R}^{m+1}$. Partition it as $\boldsymbol{\nu} \equiv (\boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$ with $\boldsymbol{\nu}_1 \in \mathbb{R}^m$. Let $\Gamma_{j,\boldsymbol{\gamma}}$ be our amplified version of the matrix Γ_j from (24), so

$$\Gamma_{j,\boldsymbol{\gamma}} \;=\; \begin{pmatrix} \Gamma_{11,j} + \operatorname{diag}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m) & \Gamma_{12,j} \\ \Gamma_{12,j}^\top & \Gamma_{22,j} \end{pmatrix}.$$

As Γ_j itself is positive semidefinite, we find that if one of the first m entries of $\boldsymbol{\nu}$ is nonzero then

$$\boldsymbol{\nu}^\top \Gamma_{j,\boldsymbol{\gamma}} \boldsymbol{\nu} \;\geq\; \boldsymbol{\nu}^\top \Gamma_j \boldsymbol{\nu} + \sum_{k=1}^{m} \boldsymbol{\nu}_k^2 \boldsymbol{\gamma}_k \;\geq\; \sum_{k=1}^{m} \boldsymbol{\nu}_k^2 \boldsymbol{\gamma}_k \;>\; 0.$$

If only the last entry of $\boldsymbol{\nu}$ is nonzero then

$$\boldsymbol{\nu}^\top \Gamma_{j,\boldsymbol{\gamma}} \boldsymbol{\nu} \;=\; \boldsymbol{\nu}_{m+1}^2 \Gamma_{22,j} \;>\; 0$$

almost surely; recall that $\boldsymbol{\Gamma}_{22,j} = \frac{1}{n}\sum_{i=1}^n h_j(X_j^{(i)})$. We conclude that $\Gamma_{j,\boldsymbol{\gamma}}$ is a.s. positive definite, which ensures unique existence of the loss minimizer.

Our methodology readily accommodates two different choices of the penalty parameter λ for $\mathbf{K}$ and $\boldsymbol{\eta}$. If the ratio of the respective values $\lambda_{\mathbf{K}}$ and $\lambda_{\boldsymbol{\eta}}$ is fixed, the proof of the above theorem can be easily modified by replacing $\boldsymbol{\eta}$ by $(\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}})\boldsymbol{\eta}$. To avoid picking two tuning parameters, one may also choose to remove the penalty on $\boldsymbol{\eta}$ altogether by profiling out $\boldsymbol{\eta}$ and solve for $\hat{\boldsymbol{\eta}} \equiv \boldsymbol{\Gamma}_{22}^{-1}\left(\boldsymbol{g}_2 - \boldsymbol{\Gamma}_{12}^\top \operatorname{vec}(\hat{\mathbf{K}})\right)$, with $\hat{\mathbf{K}}$ the minimizer of the profiled loss

$$\hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma},\text{profile}}(\mathbf{K}) \equiv \frac{1}{2}\operatorname{vec}(\mathbf{K})^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma},11.2} \operatorname{vec}(\mathbf{K}) - (\boldsymbol{g}_1 - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{g}_2)^\top \operatorname{vec}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \qquad (28)$$

where the Schur complement $\boldsymbol{\Gamma}_{\boldsymbol{\gamma},11.2} \equiv \boldsymbol{\Gamma}_{\boldsymbol{\gamma},11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{12}^\top$ is a.s. positive definite such that the profiled estimator exists a.s. for all $\lambda \geq 0$. This profiled approach corresponds to choosing $\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}} = 0$. A detailed theoretical analysis of the profiled estimator is beyond the scope of this paper. We note that in the other extreme, with $\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}} = +\infty$, the non-centered estimator reduces to the estimator from the centered case.

Remark 17 The quantity $c_{\boldsymbol{X}}$ in Theorem 16 depends on $\mathbb{E}_0 X_j$, which in turn depends on the structure of both $\boldsymbol{\mu}_0$ and $\mathbf{K}_0$. If $\mu_{0,j}$ is large compared to $(\mathbf{K}_0)_{jj}^{-1}$, then $c_{\boldsymbol{X}}$ seems to scale as $\boldsymbol{\mu}_0$, which negatively impacts the guarantees stated in Theorem 16. However, as in the one-dimensional case for estimation of μ_0 (Section 3.1), our estimator should automatically adapt to the large mean parameter. This suggests that it might be possible to improve our analysis involving $c_{\boldsymbol{X}}$.

5.3 Tuning Parameter Selection

By treating the unpenalized loss (i.e., $\lambda = 0$, $\boldsymbol{\gamma} = 0$) with the estimate plugged in as the mean negative log-likelihood, we may use the extended Bayesian Information Criterion

(eBIC) to choose the tuning parameter (Chen and Chen, 2008; Foygel and Drton, 2010). Consider the centered case as an example. Let $\hat{S}^\lambda \equiv \{(i,j) : \hat{\kappa}_{ij}^\lambda \neq 0, i < j\}$, where $\hat{\mathbf{K}}^\lambda$ be the estimate associated with tuning parameter λ. The eBIC is then

$$\text{eBIC}(\lambda) = -n\text{vec}(\hat{\mathbf{K}})^\top \boldsymbol{\Gamma}(\mathbf{x})\text{vec}(\hat{\mathbf{K}}) + 2n\boldsymbol{g}(\mathbf{x})^\top \text{vec}(\hat{\mathbf{K}}) + |\hat{S}^\lambda| \log n + 2\log \binom{p(p-1)/2}{|\hat{S}^\lambda|},$$

where $\hat{\mathbf{K}}$ can be either the original estimate associated with λ, or a refitted solution obtained by restricting the support to $\hat{S}^\lambda$.

6. Simulation Results

In this section, we present simulation results for application of our estimator to truncated GGMs with different choices of $\boldsymbol{h}$, in comparison with other competing methods.

6.1 Implementation

We optimize our loss functions with respect to a symmetric matrix $\mathbf{K}$ and in the non-centered case also the vector $\boldsymbol{\eta}$. We use a coordinate-descent method analogous to Algorithm 2 in Lin et al. (2016), where in each step we update each element of $\mathbf{K}$ and $\boldsymbol{\eta}$ based on the other entries from the previous steps, while maintaining symmetry. Warm starts at the previously considered value of the tuning parameter λ, as well as lasso-type strong screening rules (Tibshirani et al., 2012) are used for speedups. In our simulations below we always scaled the data matrix by column ℓ_2 norms before proceeding to estimation.

6.2 Choice of $\boldsymbol{h}$

Our estimator requires choosing a function $\boldsymbol{h} : \mathbb{R}_+^m \to \mathbb{R}_+^m$. We will always specify $\boldsymbol{h}(\boldsymbol{x}) = (h(x_1), \ldots, h(x_m))$ for a single non-decreasing univariate function $h : \mathbb{R}_+ \to \mathbb{R}_+$.

As previously explained, $\boldsymbol{h} \in \mathcal{H}_{M,M'}$ is a sufficient condition for assumptions (A1)-(A2), as well as (C1)-(C2) in the case of unregularized estimators. For $\boldsymbol{h} \in \mathcal{H}_{M,M'}$ to hold, h has to be bounded which we used in the proofs of our theoretical guarantees in Section 4.1. However, it is interesting to also explore other functions h that grow sub-exponentially. In particular, we will experiment with functions like $h(x) = \log(1+x)$ and $h(x) = x$. In addition, we try MCP- (Fan and Li, 2001) and SCAD-like (Zhang, 2010) functions. The results we report here are based on selections of best performing choices.

According to the subsequent simulation results, as well as additional experiments not shown here, there is no one uniformly best choice of h. We also did not observe any clear relationship between features such as convexity or the slope of h at the origin and performance of the estimator. Nonetheless, for many choices of rather simple functions h, our estimator provides a significant improvement over existing methods.

6.2.1 Truncated Centered GGMs

For data from a truncated centered Gaussian distribution, we compared our generalized score matching estimator with various choices of h, to *SpaCE JAM* (SJ, Voorman et al., 2013), which estimates graphs using additive models for conditional means, a pseudo-likelihood method *SPACE* (Peng et al., 2009) in the reformulation of Khare et al. (2015),

graphical lasso (GLASSO, Yuan and Lin, 2007; Friedman et al., 2008), the *neighborhood selection* estimator (NS) of Meinshausen and Bühlmann (2006), and finally *nonparanormal SKEPTIC* (Liu et al., 2012) with Kendall's τ. Among these competitors, only the top performing methods are explicitly shown in our reported results. Recall that the choice of $h(x) = x^2$ corresponds to the estimator from Lin et al. (2016).

For our experiment, we consider $m = 100$ variables at sample sizes $n = 80$ and $n = 1000$. For $n = 80$, amplification is necessary and based on Theorem 15 we choose the multiplier $c = C(n, m) = 1.8647$. For $n = 1000$, we consider $c = 1$, i.e., no amplification, and $c = C(n, m) = 1.6438$. The underlying inverse covariance matrices are selected as follows:

Subgraphs: Proceeding as in Section 4.2 of Lin et al. (2016), the graph is chosen to have 10 disconnected subgraphs, each containing $m/10$ nodes. Then the true parameter matrix $\mathbf{K}_0$ is block-diagonal, and in each block we generate each lower-triangular element to be 0 with probability $1-\pi$, and otherwise draw from a uniform distribution on interval $[0.5, 1]$, for some $\pi \in (0, 1)$. The upper triangular elements are determined by symmetry. The diagonal elements of $\mathbf{K}_0$ are chosen as a common positive value such that the minimum eigenvalue of $\mathbf{K}_0$ is 0.1. This setting is denoted as *SUB*.

Erdős-Rényi: For a choice of $\pi \in (0, 1)$, we generate a graph by independently including each possible edge with probability π. Non-zero entries in the lower-trianagular part of the matrix $\mathbf{K}_0$ are independent draws from the uniform distribution on $[0.5, 1]$. The matrix is then symmetrized, and the diagonal elements are set such that $\mathbf{K}_0$ has minimum eigenvalue 0.1. This type of graph is denoted as *ER*.

In each scenario, SUB and ER, we generate 5 different true precision matrices $\mathbf{K}_0$, and run 10 trials with each of these precision matrices. For $n = 1000$, we choose $\pi = 0.8$ for SUB, which is in accordance with Lin et al. (2016), and we take $\pi = 0.08$ for ER. For $n = 80$, we set $\pi = 0.2$ for SUB and $\pi = 0.02$ for ER. This way $n/(d^2_{\mathbf{K}_0} \log m)$ is roughly constant; recall Theorems 15 and 16.

We report the simulation results for setting SUB in this section, but defer the results for setting ER, which ultimately give similar conclusions, to Appendix B. For SUB, we plot representatives of the resulting ROC (*receiver operating characteristic*) curves in Figure 3, and summarize the AUCs (*areas under the curves*) for all estimators in Table 1, with their means and standard deviations over 50 curves listed. Each plotted curve corresponds to the average of 50 ROC curves, where the averaging method is the vertical averaging from Algorithm 3 in Fawcett (2006). The averaging is mean AUC-preserving.

Looking at the mean AUCs with the standard deviations in mind, all of the alternative functions h we considered perform better than $h(x) = x^2$ from Hyvärinen (2007) and Lin et al. (2016) and the competing methods. The results for $n = 1000$ in Table 1 also show that the multiplier does help improve the AUCs, a matter to be discussed in Section 6.3.

6.2.2 Truncated Non-Centered GGMs

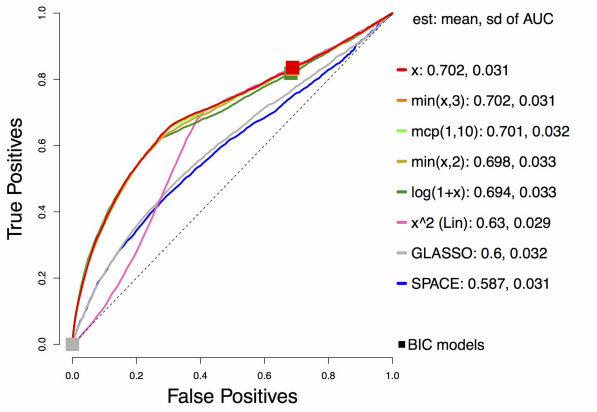
Next we generate data from a truncated non-centered Gaussian distribution with both parameters $\boldsymbol{\mu}$ and $\mathbf{K}$ unknown. In each trial we form the true $\mathbf{K}_0$ as in Section 6.2.1, and generate each component of $\boldsymbol{\mu}_0$ independently from the normal distribution with mean 0 and standard deviation 0.5.

Centered, $n = 80$, multiplier 1.8647, SUB					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.694	0.033	∞	0.702	0.031
2	0.694	0.033	3	0.702	0.031
1	0.692	0.033	2	0.698	0.033
0.5	0.664	0.038	1	0.686	0.030
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.701	0.032	10	0.702	0.031
5	0.700	0.032	5	0.701	0.032
1	0.672	0.036	2	0.696	0.033
$x^{1.5}$: (0.683, 0.030)			x^2: (0.630, 0.029)		
GLASSO (0.600,0.032)			SPACE: (0.587, 0.031)		
NS: (0.587,0.031)			SJ: (0.540,0.036)		

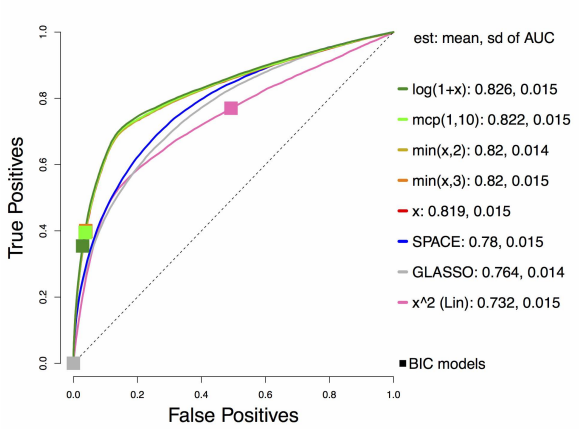
Centered, $n = 1000$, multiplier 1, SUB					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
2	0.826	0.015	2	0.820	0.014
∞	0.826	0.015	3	0.820	0.015
1	0.824	0.014	∞	0.819	0.015
0.5	0.804	0.015	1	0.817	0.014
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.824	0.015	2	0.823	0.014
10	0.822	0.015	5	0.822	0.015
1	0.810	0.015	10	0.821	0.015
$x^{1.5}$: (0.782,0.014)			x^2: (0.732,0.015)		
SPACE: (0.780,0.015)			NS: (0.779,0.015)		
GLASSO (0.764,0.014)			SJ: (0.703,0.015)		

Centered, $n = 1000$, multiplier 1.6438, SUB					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.857	0.011	3	0.855	0.011
2	0.857	0.011	∞	0.855	0.011
1	0.855	0.011	2	0.854	0.011
0.5	0.833	0.012	1	0.847	0.011
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.857	0.011	5	0.856	0.011
10	0.856	0.011	10	0.855	0.011
1	0.840	0.012	2	0.855	0.011
$x^{1.5}$: (0.812,0.011)			x^2: (0.736,0.011)		
SPACE: (0.780,0.015)			NS: (0.779,0.015)		
GLASSO (0.764,0.014)			SJ: (0.703,0.015)		

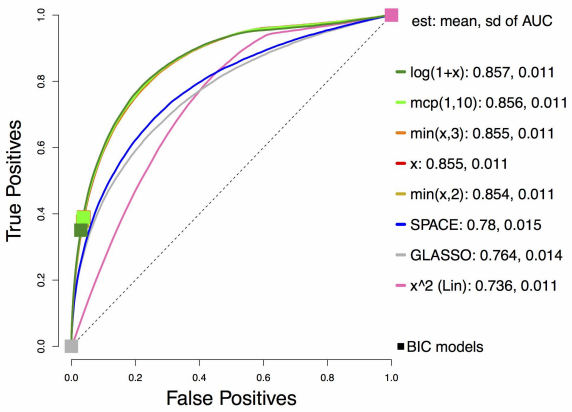
Table 1: Mean and standard deviation of areas under the ROC curves (AUC) using different estimators in the centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multiplier 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).



(a) $n = 80$, mult $= 1.8647$, SUB



(b) $n = 1000$, mult $= 1$, SUB



(c) $n = 1000$, mult $= 1.6438$, SUB

Figure 3: Average ROC curves of our *centered* estimator for $m = 100$ variables and two sample sizes n under various choices of h, compared to SPACE and GLASSO, for the *truncated centered GGM* case. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.

Non-centered profiled, $n = 80$, multiplier 1.8647, SUB					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.632	0.032	∞	0.634	0.032
2	0.632	0.032	3	0.634	0.032
1	0.631	0.032	2	0.632	0.032
0.5	0.619	0.033	1	0.628	0.032
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.634	0.032	5	0.634	0.032
5	0.634	0.032	10	0.634	0.032
1	0.622	0.032	2	0.634	0.032
$x^{1.5}$: (0.623,0.031)			x^2: (0.607,0.030)		
GLASSO: (0.614,0.029)			NS: (0.604,0.028)		
SPACE: (0.602,0.029)			SJ: (0.561,0.036)		

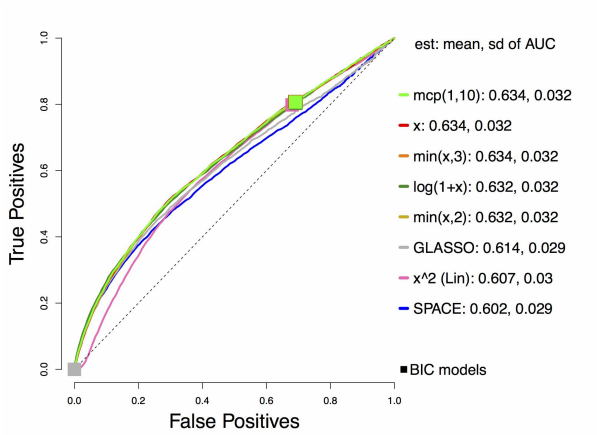
Non-centered profiled, $n = 1000$, multiplier 1, SUB					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.783	0.020	2	0.779	0.020
2	0.783	0.020	∞	0.779	0.020
1	0.782	0.020	3	0.779	0.020
0.5	0.767	0.021	0.5	0.758	0.020
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.782	0.020	2	0.780	0.020
10	0.780	0.020	5	0.780	0.020
1	0.771	0.021	10	0.779	0.020
$x^{1.5}$: (0.751,0.019)			x^2: (0.713,0.018)		
SPACE: (0.786,0.020)			NS: (0.785,0.02)		
GLASSO (0.770,0.019)			SJ: (0.720,0.019)		

Non-centered profiled, $n = 1000$, multiplier 1.6438, SUB					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.764	0.018	∞	0.766	0.019
2	0.764	0.018	3	0.765	0.019
1	0.762	0.018	2	0.764	0.018
0.5	0.738	0.018	1	0.753	0.018
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.766	0.019	10	0.766	0.019
5	0.766	0.019	5	0.766	0.019
1	0.745	0.018	2	0.763	0.018
$x^{1.5}$: (0.748,0.018)			x^2: (0.718,0.017)		
SPACE: (0.786,0.020)			NS: (0.785,0.020)		
GLASSO (0.770,0.019)			SJ: (0.720,0.019)		

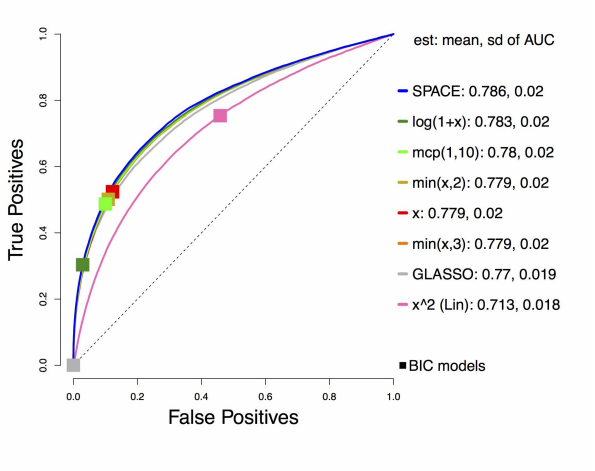
Table 2: Mean and standard deviation of AUC using different *profiled* estimators in the non-centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multipliers 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).

We report results for our *profiled* estimator based on (28), with no penalty on $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$, with different h, and make a comparison with SPACE, SpaCE JAM (SJ), GLASSO, and neighborhood selection (NS), each with 50 trials as before. As before, representative ROC curves are plotted in Figure 4, with AUCs for all estimators summarized in Table 2.

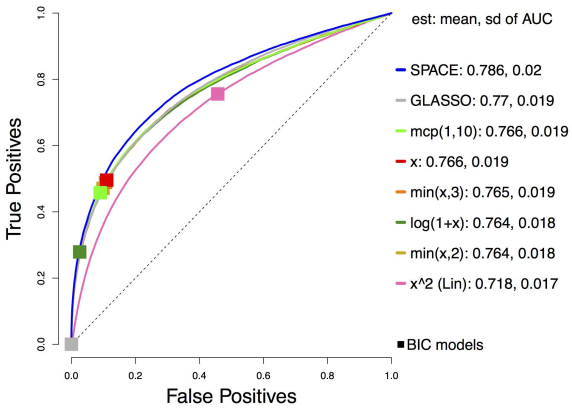
Even without tuning the extra penalty parameter on $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$, our profiled estimator beats the competing methods by a large margin when $n = 80$. For the SUB graph, with multipliers 1 and $n = 1000$ our estimators still do better than Space JAM and GLASSO, and have performance comparable to other competing methods. It might appear that the performance of our estimators deterioriate with a multiplier larger than 1, but as we will



(a) $n = 80$, mult $= 1.8647$, SUB

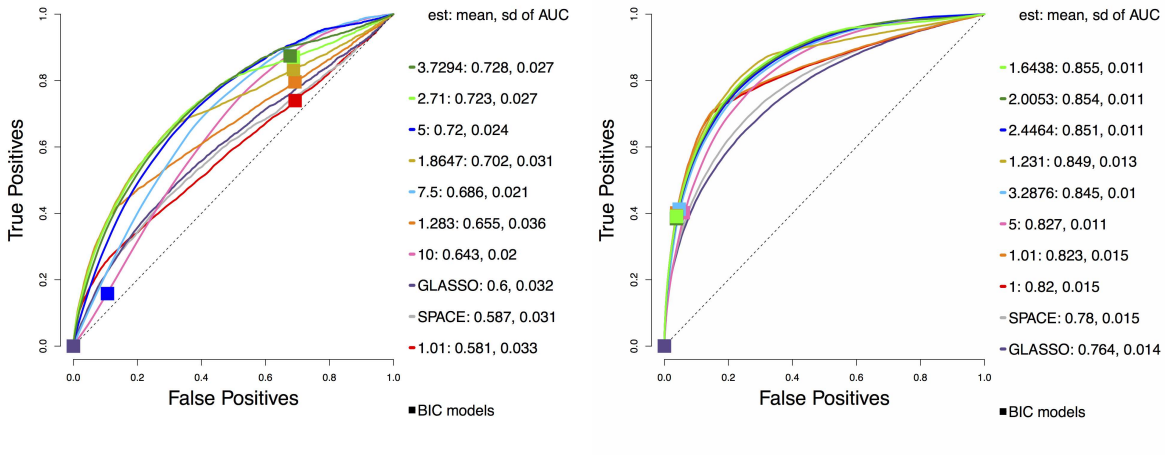


(b) $n = 1000$, mult $= 1$, SUB



(c) $n = 1000$, mult $= 1.6438$, SUB

Figure 4: Average ROC curves of our *non-centered profiled* estimator with various choices of h, compared to SPACE and GLASSO, for the *truncated non-centered GGM* case. $n = 80$ or 1000, $m = 100$. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.



(a) $n = 80$, SUB

(b) $n = 1000$, SUB

Figure 5: Performance of $\min(x, 3)$ for truncated centered GGMs using different multipliers, compared to GLASSO and SPACE, in the centered setting, $n = 80$ or 1000.

see there can be significant improvement in the AUC if we tune an additional parameter, in the presence of a multiplier larger than 1.

As in the centered case, the leading h functions in each category perform similarly, and the exact choice is not crucial. Subsequently, we will simply use $h(x) = \min(x, 3)$.

6.3 Choice of multiplier

6.3.1 Truncated Centered GGMs

In Figure 5 we consider the centered case as in Section 6.2.1 and compare the ROC curves for GLASSO, SPACE, and $h(x) = \min(x, 3)$ with different levels of amplification specified via different multipliers c. Theorem 15 guarantees consistency only for $c < C(n, m)$ but we observe that there can be a gain from going beyond the *upper-bound multiplier* $C(n, m)$, which is 1.8647 for $n = 80$ and 1.6438 for $n = 1000$ (in the $n = 1000$ simulation $C(n, m)$ turns out to be the best-performing multiplier). However, the effect deteriorates fast as the multiplier grows larger. The figure suggests that while some additional gains are possible by tuning over the choice of multiplier, the *upper-bound multiplier* is a good default.

6.3.2 Truncated Non-Centered GGMs

In Figure 6, we take up the simulation setup for the non-centered case from Section 6.2.2, and use the non-profiled estimator, that is, the non-centered estimator with ℓ_1 penalty on both $\mathbf{K}$ and $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$. The ROC curves are compared to competing methods GLASSO and SPACE. For amplification underlying our estimator we consider the upper-bound multiplier $C(n, m)$ from Theorem 16 as one default. We refer to this as *high* amplification. We also consider certain lower amount of amplification, with $c = 2 - (1 + 24e \log m/n)^{-1}$, referred to

as *medium*. Finally, for $n = 1000$ we also consider a *low* multiplier 1, which corresponds to no amplification. We compare these possible defaults to a finer grid of multipliers of which we show some representatives in the plots.

We see that among our defaults the upper-bound choice $C(n, m)$ performs best. Some additional gains are possible by tuning the multiplier over a grid of values containing this choice. Moreover, we see that it can be beneficial to tune over both $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$.

We remark that while for one run, the best model picked by BIC falls on the ROC curve, in (c) of Figure 6 a few squares are off the curve since these squares correspond to the average of the true and false positive rates of the chosen BIC models over all 50 runs. But in all cases, the average of the models picked by BIC tuned over both $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$ looks reasonable.

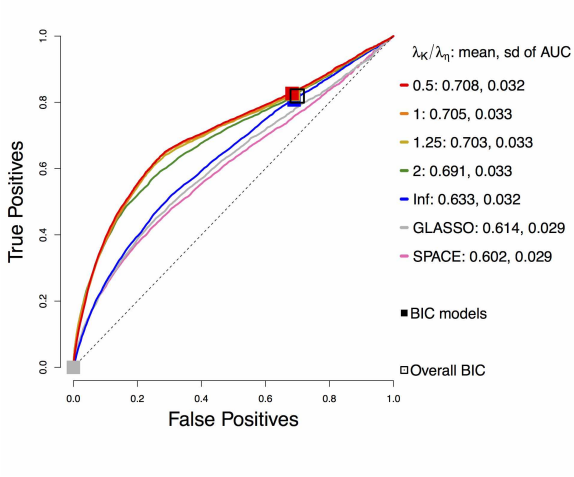
6.4 RNAseq Data

In this section we apply our regularized generalized $\boldsymbol{h}$-score matching estimator for truncated non-centered GGMs to RNAseq data also studied in Lin et al. (2016). We have $n = 487$ prostate adenocarcinoma samples from The Cancer Genome Atlas (TCGA) dataset. Following Lin et al. (2016), we focus on $m = 333$ genes that belong to the known cancer pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) and that have no more than 10% missing values. Missing values are set to 0. We choose $h(x) = \min(x, 3)$ and use the upper-bound multiplier (*high*), as discussed in Section 6.3. For simplicity, we use the profiled estimator, and choose the regularization parameter λ so that the estimated graph has exactly $m = 333$ edges, all these choices being as in Lin et al. (2016) whose estimator corresponds to $h(x) = x^2$ without amplification.

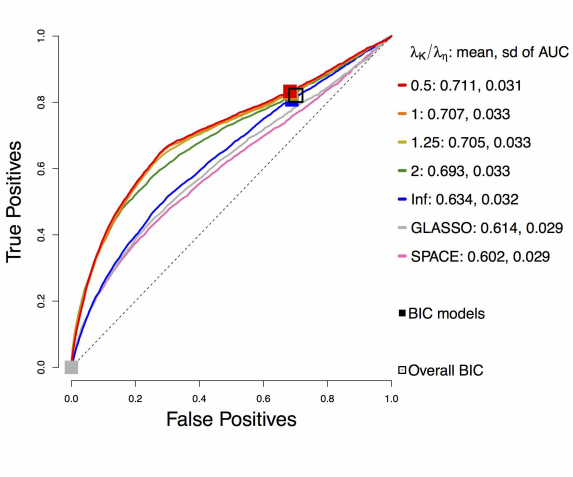
We compare our graph to the one in Lin et al. (2016), which corresponds to $h(x) = x^2$ with no multiplier. Shown in Figure 7 are the estimated graphs, and their intersection graph in the middle, with isolated nodes removed and layout optimized for each plot. Red-colored points are the "hub nodes", namely nodes with degree at least 10. In Figure 8 we plot the same graphs in a fixed layout optimized for the graph corresponding to $h = \min(x, 3)$, and include the isolated nodes.

Out of 333 edges, the two estimated graphs share 117 edges in common. Assuming edges are placed at random between nodes and the two graphs are independent, the distribution of the number R of common edges follow a hypergeometric distribution $P(R = r) = \frac{\binom{m}{r}\binom{m(m-1)/2-m}{m-r}}{\binom{m(m-1)/2}{m}}$; for $m = 333$ the probability that we get at least 117 common edges is essentially 0. Thus, the large number of shared edges between the two methods can be explained by the fact that they both minimize the same underlying score-matching loss.

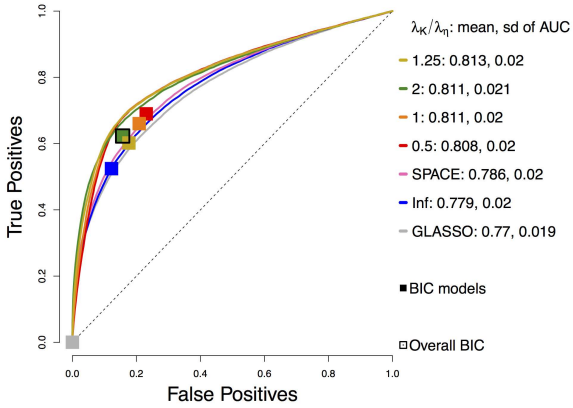
Figure 9 shows histograms (boxplots) of the node degree distribution for each graph, with common y-limits. It is obvious that the graph using $h(x) = \min(x, 3)$ has much more isolated nodes than the other, and has a slightly smaller max degree. As in Lin et al. (2016), using the Kolmogorov-Smirnov test we also fail to reject the null hypothesis that the node degrees follow a power law distribution at any practical significance level. Table 3 provides another way of comparing between the two graphs by listing the genes with the highest node degrees.



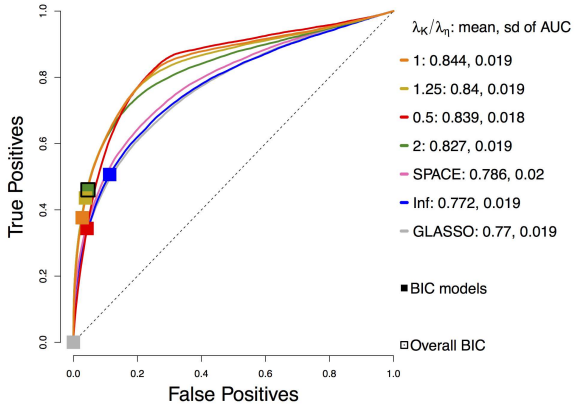
(a) $n = 80$, mult $= 1.7897$, SUB



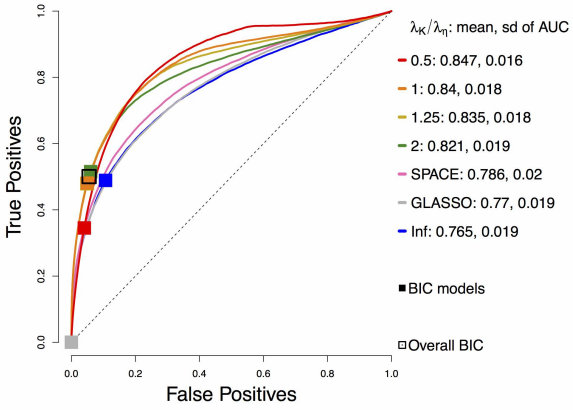
(b) $n = 80$, mult $= 1.8647$, SUB



(c) $n = 1000$, mult $= 1$, SUB

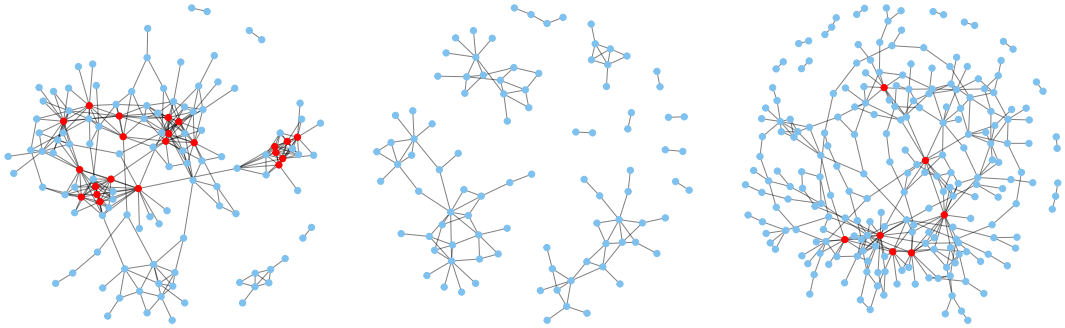


(d) $n = 1000$, mult $= 1.2310$, SUB



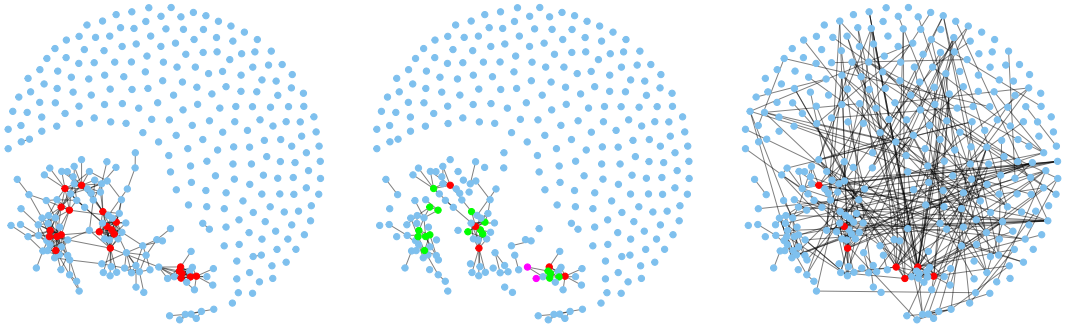
(e) $n = 1000$, mult $= 1.6438$, SUB

Figure 6: Performance of the non-centered estimator with $h(x) = \min(x, 3)$. Each curve corresponds to a different choice of $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$. Squares indicate models picked by eBIC with refit. The square with black outline has highest eBIC among all models (combinations of $\lambda_{\mathbf{K}}$, $\lambda\boldsymbol{\eta}$). Multipliers correspond to medium or high for $n = 80$, and low, medium or high for $n = 1000$, respectively.



(a) $\min(x, 3)$ (b) Common edges (c) Lin et al.

Figure 7: Graphs estimated by regularized generalized score matching estimator with $h(x) = \min(x, 3)$ with upper-bound multiplier (left) and $h(x) = x^2$ with no multiplier (Lin et al., 2016, right), and their intersection graph (middle). Isolated nodes with no edges are removed, and the layout is optimized for each plot. In (a) and (c), red points indicate nodes with degree at least 10 ("hub nodes").



(a) $\min(x, 3)$ (b) Common edges (c) Lin et al.

Figure 8: Graphs estimated by regularized generalized score matching estimator with $h(x) = \min(x, 3)$ with upper-bound multiplier (left) and $h(x) = x^2$ with no multiplier (Lin et al., 2016, right), and their intersection graph (middle). Isolated nodes are included and the layout is fixed across plots and optimized for graph (a). In (b) the red nodes are hub nodes shared by both graphs, the green ones are hub nodes in graph (a) only. Magenta nodes are hub nodes in graph (c) only.

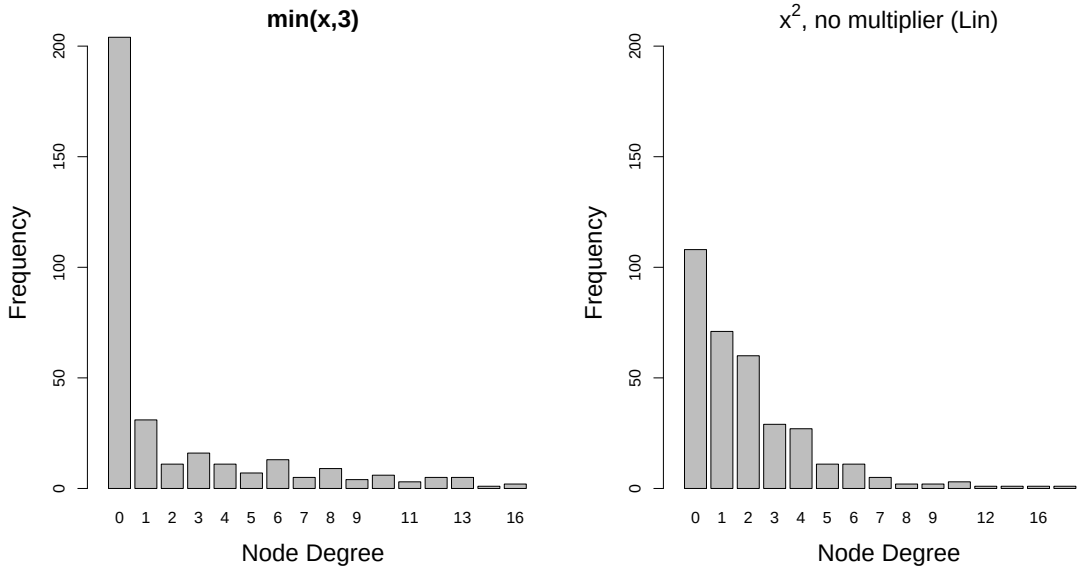


Figure 9: Node degree distributions for the two estimated graphs.

$\min(x, 3)$ with multiplier 1.63	Lin et al.
LAMB3 (16)	CCNE2 (19)
PIK3CG (16)	**PIK3CG (16)**
MMP2 (15)	BRCA2 (13)
GLI2 (13)	**BIRC5 (12)**
LAMA4 (13)	**LAMB3 (10)**
PDGFRB (13)	**PIK3CD (10)**
PIK3CD (13)	SKP2 (10)
RASSF5 (13)	HRAS (9)
BIRC5 (12)	STAT5B (9)
FLT3 (12)	**GSTP1(8)**
GSTP1 (12)	**PDGFRB (8)**
LAMA2 (12)	
RAC2 (12)	

Table 3: List of genes with the highest node degrees in each estimated graph.

In Table 3 we list the top ten genes in terms of node degree for both estimated graphs. Due to ties, 13 genes are listed for $h(x) = \min(x, 3)$ and 11 for Lin et al. (2016). Among these top genes, six are common in both graphs, and are discussed in Lin et al. (2016). We next elaborate on the evidence supporting the first four of the newly discovered genes.

- MMP2 (Matrix metalloproteinase 2): According to Trudel et al. (2003), increased MMP-2 expression is an independent predictor of decreased prostate cancer disease-free survival. Morgia et al. (2005) state that the activity of MMP-2 can be useful in diagnosis, therapy, and assessment of malignant cancer progression. Other references include Stearns and Stearns (1996), Wilson et al. (2004) and Xie et al. (2016).

- GLI2 (GLI family zinc finger 2): GLI2 is a primary mediator of the hedgehog signaling pathway, which has been reported in prostate cancer, and plays a critical role in the malignant phenotype of prostate cancer cells (Thiyagarajan et al., 2007). Its increased level of expression is also related to AI prostate cancer, and may be a therapeutic target in castrate-resistant prostate cancer (Narita et al., 2008).

- LAMA4 (Laminin subunit alpha 4): LAMA4 is consistently upregulated in benign prostatic hyperplasia when compared to normal prostate tissues (Luo et al., 2002).

- RASSF5 (RAS association domain family member 5): The combination of RASSF5 along with four other DNA methylation markers can effectively differentiate between benign prostate biopsy cores from non-cancer patients and cancer cores, and can be used to identify patients at risk without repeat biopsies (Brikun et al., 2014).

7. Discussion

In this paper we proposed a generalized version of the score matching estimator of Hyvärinen (2007), which avoids the calculation of normalizing constants. For estimation of the canonical parameters of exponential families, our generalized loss retains the nice property of being quadratic in the parameters. Our estimator offers improved estimation properties through various scalar or vector-valued choices of function $\boldsymbol{h}$.

For high-dimensional exponential family graphical models, following the work of Meinshausen and Bühlmann (2006), Yuan and Lin (2007) and Lin et al. (2016), we add an ℓ_1 penalty to the generalized score matching loss, giving a solution that is almost surely unique under regularity conditions and has a piecewise linear solution path.

In the case of multivariate truncated Gaussian distribution, where the conditional independence graph is given by the inverse covariance parameter, the sample size required for the consistency of our method is $\Omega(d^2 \log m)$, where m is the dimension and d is the maximum node degree in the corresponding independence graph. This matches the rates for Gaussian graphical models in Ravikumar et al. (2011) and Lin et al. (2016), and lasso with linear regression (Bühlmann and van de Geer, 2011).

One issue that is overlooked in Lin et al. (2016) and related earlier work is the fact that the score matching loss can be unbounded below for a small tuning parameter, when $m > n$. We fix this issue by amplifying the diagonal entries in the quadratic matrix in the definition of the generalized score matching loss by a factor/multiplier, and give an upper bound on

that multiplier that guarantees consistency. Introducing this multiplier also results in a substantial gain in performance.

A potential problem for future work would be adaptive choice of $\boldsymbol{h}$ from data, or to develop a summary score similar to eBIC that can be used to compare not just different tuning parameters but also across different models.

8. Proofs

The following lemma is used in the proof of Theorem 4.

Lemma 18 *Assuming that f and g are differentiable a.e., then for all $j = 1, \dots, m$,*

$$\lim_{a \nearrow +\infty,\, b \searrow 0^+} f(\boldsymbol{x}_{-j}; a) g(\boldsymbol{x}_{-j}; a) - f(\boldsymbol{x}_{-j}; b) g(\boldsymbol{x}_{-j}; b) = \int_0^\infty f(\boldsymbol{x}) \frac{\partial g(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}x_j + \int_0^\infty g(\boldsymbol{x}) \frac{\partial f(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}x_j,$$

where $(\boldsymbol{x}_{-j}; a)$ is the vector obtained by replacing the j-th component of $\boldsymbol{x}$ by a.

Proof This is in analogy of Lemma 4 from Hyvärinen (2005), which is proved by integrating the partial derivatives. ■

Proof [Proof of Theorem 4] Recall the following assumptions from Section 2.2:

(A1) $p_0(\boldsymbol{x}) h_j(x_j) \partial_j \log p(\boldsymbol{x}) \to 0$ as $x_j \nearrow +\infty$ and as $x_j \searrow 0^+$ $\forall \boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}$ $\forall p \in \mathcal{P}_+$,

(A2) $\mathbb{E}_{p_0} \|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 < +\infty$, $\mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 < +\infty$ $\forall p \in \mathcal{P}_+$,

where

$$\partial_j \log p(\boldsymbol{x}) \equiv \left. \frac{\partial \log p(\boldsymbol{y})}{\partial y_j} \right|_{\boldsymbol{y} = \boldsymbol{x}}.$$

Without explicitly stating the domains $\mathbb{R}_+$ or $\mathbb{R}_+^m$ in the integrals in the equations, the function $J_{\boldsymbol{h}}(p)$ from (5) is equal to

$$\begin{aligned}
&\frac{1}{2} \int p_0(\boldsymbol{x}) \Big[\|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})\|_2^2 - 2 (\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x}))^\top (\nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})) \\
&\qquad + \|\nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})\|_2^2 \Big] \,\mathrm{d}\boldsymbol{x} \\
&= \underbrace{\frac{1}{2} \int p_0(\boldsymbol{x}) \sum_{j=1}^m h_j(x_j) \left(\frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \right)^2 \mathrm{d}\boldsymbol{x}}_{\equiv A} \underbrace{- \int p_0(\boldsymbol{x}) \sum_{j=1}^m h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}\boldsymbol{x}}_{\equiv B} \\
&\qquad \underbrace{+ \frac{1}{2} \int p_0(\boldsymbol{x}) \sum_{j=1}^m h_j(x_j) \left(\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \right)^2 \mathrm{d}\boldsymbol{x}}_{\equiv C}.
\end{aligned}$$

Here, A is already in the desired form of an integral with respect to $p_0(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}$ whose integrand does not depend on p_0. The third term C is a constant in the sense that it only involves

the true density p_0. It remains to simplify B by integration by parts. Note also that we can split the integral into these three parts because A and C are assumed finite in the first part of (A2), and the integrand in B is integrable since $|2ab| \leq a^2 + b^2$.

By linearity and Fubini's theorem, we can write

$$\begin{aligned} B &= -\sum_{j=1}^{m} \int p_0(\boldsymbol{x}) h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}\boldsymbol{x} \\ &= -\sum_{j=1}^{m} \int \left[\int p_0(\boldsymbol{x}) h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}. \end{aligned}$$

As $\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} = \frac{1}{p_0(\boldsymbol{x})} \frac{\partial p_0(\boldsymbol{x})}{\partial x_j}$, this can be simplified to

$$B = -\sum_{j=1}^{m} \int \left[\int \frac{\partial p_0(\boldsymbol{x})}{\partial x_j} h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \,\mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}.$$

Then by Lemma 18 and assumption (A1),

$$\begin{aligned} B = -\sum_{j=1}^{m} \int \Bigg[& \lim_{a \nearrow \infty, b \searrow 0^+} \left[p_0(\boldsymbol{x}_{-j}; a) h_j(a) \partial_j \log p(\boldsymbol{x}_{-j}, a) - p_0(\boldsymbol{x}_{-j}; b) h_j(b) \partial_j \log p(\boldsymbol{x}_{-j}, b) \right] \\ & - \int p_0(\boldsymbol{x}) \frac{\partial \left(h_j(x_j) \partial_j \log p(\boldsymbol{x}) \right)}{\partial x_j} \,\mathrm{d}x_j \Bigg] \mathrm{d}\boldsymbol{x}_{-j} \\ = \sum_{j=1}^{m} \int & \left[\int p_0(\boldsymbol{x}) \frac{\partial (h_j(x_j) \partial_j \log p(\boldsymbol{x}))}{\partial x_j} \,\mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}. \end{aligned}$$

Justified by the second half of (A2), by Fubini-Tonelli and linearity again

$$\begin{aligned} B &= \sum_{j=1}^{m} \int p_0(\boldsymbol{x}) \frac{\partial (h_j(x_j) \partial_j \log p(\boldsymbol{x}))}{\partial x_j} \,\mathrm{d}\boldsymbol{x}, \\ &= \sum_{j=1}^{m} \int h_j'(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} p_0(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} + \sum_{j=1}^{m} \int h_j(x_j) \frac{\partial^2 \log p(\boldsymbol{x})}{\partial x_j^2} p_0(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \end{aligned}$$

Thus,

$$\begin{aligned} & J_{\boldsymbol{h}}(p) = B + A + C = \\ & \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[h_j'(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} + h_j(x_j) \frac{\partial^2 \log p(\boldsymbol{x})}{\partial x_j^2} + \frac{1}{2} h_j(x_j) \left(\frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \right)^2 \right] \mathrm{d}\boldsymbol{x} + C, \end{aligned}$$

where C does not depend on p. ■

Proof [Proof of Theorem 5] For exponential families, under the assumptions the empirical loss $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ in (7) becomes (up to an additive constant)

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[h_j'(X_j^{(i)}) \frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial X_j^{(i)}} + h_j(X_j^{(i)}) \frac{\partial^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial (X_j^{(i)})^2} + \frac{1}{2} h_j(X_j^{(i)}) \left(\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial X_j^{(i)}} \right)^2 \right]$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\Big[h_j'(X_j^{(i)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + b_j'(\boldsymbol{X}^{(i)})) + h_j(X_j^{(i)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + b_j''(\boldsymbol{X}^{(i)})) \\
&\qquad + \frac{1}{2}h_j(X_j^{(i)})(\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + b_j'(\boldsymbol{X}^{(i)}))^2\Big] \\
&= \frac{1}{n}\left\{\frac{1}{2}\boldsymbol{\theta}^\top\left[\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top\right]\boldsymbol{\theta}\right. \\
&\quad \left. + \left[\sum_{i=1}^{n} h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\right]^\top\boldsymbol{\theta}\right\} + \text{const},
\end{aligned}
$$

which is quadratic in $\boldsymbol{\theta}$. Let

$$
\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top, \tag{29}
$$

$$
\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\Big[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\Big]. \tag{30}
$$

Then $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top\boldsymbol{\theta} + \text{const}$. ■

Proof [Proof of Theorem 7] We estimate $\theta \equiv \mu/\sigma^2$. By (10) and (11),

$$
\begin{aligned}
\hat{\mu}_h &= \sigma^2\hat{\theta} \equiv \sigma^2\Gamma(\boldsymbol{x})^{-1}g(\boldsymbol{x}) \\
&= -\sigma^2\left[\sum_{i=1}^{n} h(X_i)t'(X_i)^2\right]^{-1}\left[\sum_{i=1}^{n} h(X_i)b'(X_i)t'(X_i) + h(X_i)t''(X_i) + h'(X_i)t'(X_i)\right] \\
&= -\sigma^2\left[\sum_{i=1}^{n} h(X_i)\right]^{-1}\left[\sum_{i=1}^{n} -h(X_i)\frac{X_i}{\sigma^2} + h'(X_i)\right].
\end{aligned}
$$

By Theorem 6,

$$
\begin{aligned}
\sqrt{n}(\hat{\mu}_h - \mu_0) &\to_d \mathcal{N}\left(0, \frac{\sigma^4\mathbb{E}_0\left[h(X)\frac{\mu_0 - X}{\sigma^2} + h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}\right) \\
&= \mathcal{N}\left(0, \frac{\mathbb{E}_0\left[h(X)(\mu_0 - X) + \sigma^2 h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}\right).
\end{aligned}
$$

By integration by parts, (suppressing the dependence of p_{μ_0} on μ_0)

$$
\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X - \mu_0)] \\
&= \int_0^\infty h'(x)h(x)(x - \mu_0)p(x)\,\mathrm{d}x = \int_0^\infty h(x)(x - \mu_0)p(x)\,\mathrm{d}h(x)
\end{aligned}
$$

$$= h^2(x)(x-\mu_0)p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu_0)p(x)$$
$$= -\int h^2(x)p(x)\,\mathrm{d}x - \int h(x)h'(x)(x-\mu_0)p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu_0)^2}{\sigma^2}p(x)\,\mathrm{d}x,$$

where we used the assumption that $\lim_{x\searrow 0^+} h(x) = 0$ and $\lim_{x\nearrow+\infty} h^2(x)(x-\mu_0)p_{\mu_0}(x) = 0$ in the last step. So

$$\mathbb{E}_0[h(X)h'(X)(X-\mu_0)] = \frac{\mathbb{E}[h^2(X)((X-\mu_0)^2/\sigma^2-1)]}{2}. \tag{31}$$

The asymptotic variance becomes

$$\frac{\mathbb{E}_0\left[h(X)(\mu_0-X)+\sigma^2h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}$$
$$= \frac{\mathbb{E}_0\left[h^2(X)(X-\mu_0)^2 - 2\sigma^2h^2(X)\left((X-\mu_0)^2/\sigma^2-1\right)/2 + \sigma^4h'^2(X)\right]}{\mathbb{E}_0^2[h(X)]}$$
$$= \frac{\mathbb{E}_0[\sigma^2h^2(X)+\sigma^4h'^2(X)]}{\mathbb{E}_0^2[h(X)]}.$$

The Cramér-Rao lower bound follows from taking the second derivative of $\log p_{\mu_0}$ with respect to μ_0. ■

Proof [Proof of Theorem 8] We estimate $\theta \equiv 1/\sigma^2$. By (10) and (11),

$$\hat\theta \equiv \Gamma(\boldsymbol{x})^{-1}g(\boldsymbol{x})$$
$$= -\left[\sum_{i=1}^n h(X_i)t'(X_i)^2\right]^{-1}\left[\sum_{i=1}^n h(X_i)b'(X_i)t'(X_i) + h(X_i)t''(X_i) + h'(X_i)t'(X_i)\right]$$
$$= \left[\sum_{i=1}^n h(X_i)(X_i-\mu)^2\right]^{-1}\left[\sum_{i=1}^n h(X_i) + h'(X_i)(X_i-\mu)\right].$$

By Theorem 6, $\sqrt{n}(\hat\theta-\theta) \to_d \mathcal{N}(0,\varsigma^2)$, where

$$\varsigma^2 \equiv \frac{\mathbb{E}_0\left[h(X)((X-\mu)^2/\sigma_0^2-1) - h'(X)(X-\mu)\right]^2}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}$$
$$= \frac{1}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\Big(\mathbb{E}_0[h^2(X)(X-\mu)^4/\sigma_0^4 - 2h^2(X)(X-\mu)^2/\sigma_0^2 + h^2(X)$$
$$+ h'^2(X)(X-\mu)^2 - 2h(X)h'(X)(X-\mu)^3/\sigma_0^2 + 2h(X)h'(X)(X-\mu)\Big).$$

By integration by parts, (suppressing the dependence of $p_{\sigma_0^2}$ on σ_0^2)

$$\mathbb{E}_0[h(X)h'(X)(X-\mu)^3]$$

$$= \int_0^\infty h'(x)h(x)(x-\mu)^3 p(x)\,\mathrm{d}x = \int_0^\infty h(x)(x-\mu)^3 p(x)\,\mathrm{d}h(x)$$
$$= h^2(x)(x-\mu)^3 p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu)^3 p(x)$$
$$= -\int h(x)h'(x)(x-\mu)^3 p(x)\,\mathrm{d}x - 3\int h^2(x)(x-\mu)^2 p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu)^4}{\sigma_0^2}p(x)\,\mathrm{d}x,$$

where we used the assumption that $\lim_{x\searrow 0^+} h(x) = 0$ and $\lim_{x\nearrow+\infty} h^2(x)(x-\mu)^3 p_{\sigma_0^2}(x) = 0$ in the last step. Combining this with (31) we get

$$\sqrt{n}(\hat{\theta}-\theta) \to_d \mathcal{N}(0,\varsigma^2) \sim \mathcal{N}\left(0, \frac{2\mathbb{E}_0[h^2(X)(X-\mu)^2/\sigma_0^2] + \mathbb{E}_0[h'^2(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right),$$

and so by the delta method, for $\hat{\sigma}_k^2 \equiv \hat{\theta}^{-1}$,

$$\sqrt{n}(\hat{\sigma}_h^2 - \sigma_0^2) \to_d \mathcal{N}\left(0, \frac{2\sigma_0^6\mathbb{E}_0[h^2(X)(X-\mu)^2] + \sigma_0^8\mathbb{E}_0[h'^2(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right).$$

The Cramér-Rao lower bound follows from taking the second derivative of $\log p_{\sigma_0^2}$ with respect to σ_0^2. ■

Proof [Proof of Corollary 13] By Theorem 12, under assumptions in that theorem, the support of $\hat{\mathbf{\Psi}}$ is a subset of the true support of $\mathbf{\Psi}_0$, and $\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_\infty \le \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda$. Since $\mathbf{\Psi}_0$ has $|S_0|$ nonzero entries,

$$|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_F = \left[\sum_{\mathbf{\Psi}_{0,jk}\neq 0} (\hat{\mathbf{\Psi}}_{jk} - \mathbf{\Psi}_{0,jk})^2\right]^{1/2} \le \sqrt{|S_0|}|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_\infty \le \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|}.$$

Similarly, by the definition of matrix ℓ_∞-ℓ_∞ norm,

$$|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_2 \le |||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_\infty = \max_{j=1,\ldots,m}\sum_{k=1}^m |\hat{\mathbf{\Psi}}_{jk} - \mathbf{\Psi}_{0,jk}| \le \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda d_{\mathbf{\Psi}_0}.$$

The result follows by also noting that $|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_2 \le |||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_F$. ■

Proof [Proof of Theorem 14] The proof is based on Theorem 12 and a probabilistic bound on $\|\mathbf{\Gamma}_{\boldsymbol{\gamma}} - \mathbf{\Gamma}_0\|_\infty$, where in the case of centered Gaussian $\mathbf{\Gamma} = \frac{1}{n}\mathrm{diag}(\mathbf{x}\mathbf{x}^\top, \ldots, \mathbf{x}\mathbf{x}^\top)$. Denote $\mathbf{\Sigma}_0 = \mathbf{K}_0^{-1}$. In particular, given $\tau > 2$ we wish to show that for $\epsilon = 80\sqrt{2}c_0\max_j(\Sigma_{0,jj})$ assuming $c_0 \equiv \sqrt{(\tau\log m + \log 4)/n} < 1/\sqrt{2}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_j^{(i)}X_k^{(i)} + \gamma_{1j}\mathbb{1}_{j=k} - \mathbb{E}X_jX_k\right| > \epsilon\right) \le m^{2-\tau},$$

and so the results follow from Theorem 12.

By Lemma 1 of Ravikumar et al. (2011), since $X_j/\sqrt{\Sigma_{0,jj}}$ is Gaussian with mean 0 and standard deviation 1,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} - \mathbb{E}X_j X_k\right| > t\right) \leq 4\exp\left(-\frac{nt^2}{3200\max_j(\Sigma_{0,jj})^2}\right)$$

for $t \in (0, 40\max_j(\Sigma_{0,jj}))$. Denote the event as $\mathcal{E}_{j,k}(t)$. Note that $\mathbb{E}X_j^2 \leq \max_j \Sigma_{0,jj} = \epsilon/(80\sqrt{2}c_0)$. Then letting $t = \epsilon/2$ and conditioning on $\mathcal{E}_{j,j}(\epsilon/2)^{\complement}$ we have

$$\frac{1}{n}\sum_{i=1}^{n} {X_j^{(i)}}^2 \leq \mathbb{E}X_j^2 + \epsilon/2 \leq \frac{\epsilon}{2}\left(1 + \frac{1}{40\sqrt{2}c_0}\right).$$

Thus by choosing $\gamma_{\ell j} = (c-1)\sum_{i=1}^{n} {X_j^{(i)}}^2/n$ for $\ell = 1, \ldots, m$ ($\mathbf{\Gamma}$ has m identical blocks) with $1 < c < 1 + \left(1 + 1/(40\sqrt{2}c_0)\right)^{-1}$, by triangle inequality and a union bound we have

$$\mathbb{P}\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} + \gamma_{1j}\mathbb{1}_{j=k} - \mathbb{E}X_j X_k\right| > \epsilon\right) \leq \mathbb{P}(\mathcal{E}_{j,k}(\epsilon/2)) = m^{2-\tau}.$$

Since $\tau > 2$, $1 + \left(1 + 1/(40\sqrt{2}c_0)\right)^{-1} = 2 - (1 + 40\sqrt{2}c_0)^{-1} > 2 - (1 + 80\sqrt{\log m/n})^{-1} \equiv C(n,m)$, so it is safe to choose any $c \in (1, C(n,m))$.

Hence, by the requirement on ϵ, our claim holds when $n > \max(c^* c_1^2 d_{\mathbf{K}}^2, 2)(\tau\log m + \log 4)$ with $c^* = 12800\max_j(\Sigma_{0,jj})^2$, compared to $c^* = 3200\max_j(\Sigma_{0,jj})^2$ in Lin et al. (2016). ■

Proof [Proof of Theorem 15] The proof of Theorem 12 from Lin et al. (2016) does not rely on the fact that the original $\mathbf{\Gamma}$ is an unbiased estimator for the population $\mathbf{\Gamma}_0$, but instead only requires one to bound $\|\mathbf{\Gamma} - \mathbf{\Gamma}_0\|_\infty$. Thus, for $\mathbf{\Gamma}_{\boldsymbol{\gamma}} = \mathbf{\Gamma} + \mathrm{diag}(\boldsymbol{\gamma})$, by Theorem 12 it suffices to prove that for any $\tau > 3$, we can bound $\|\mathbf{\Gamma}(\mathbf{x}) + \mathrm{diag}(\boldsymbol{\gamma}(\mathbf{x})) - \mathbf{\Gamma}_0\|_\infty$ by some ϵ_1 and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty$ by some ϵ_2, uniformly with probability $1 - m^{3-\tau}$. Recall from (19) that the j^{th} block of $\mathbf{\Gamma}_{\boldsymbol{\gamma}} \in \mathbb{R}^{m^2 \times m^2}$ has (k,ℓ)-th entry

$$\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) + \gamma_{kj} \cdot \mathbb{1}_{k=\ell}.$$

The entry in $\boldsymbol{g} \in \mathbb{R}^{m^2}$ (obtained by linearizing a $m \times m$ matrix) corresponding to (j,k) is

$$\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j'(X_j^{(i)}) + \mathbb{1}_{j=k} \cdot \frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}).$$

Denote $M \equiv \max_j \sup_{x>0} h_j(x)$ and $M' \equiv \max_j \sup_{x>0} h_j'(x)$, and let $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}_0 X_j)$. Using results for sub-gaussian random variables from Lemma 22.2, we have for any $t_1 > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| > t_1\right) \leq 2\exp\left(-\min\left(\frac{nt_1^2}{2M^2 c_{\boldsymbol{X}}^4}, \frac{nt_1}{2Mc_{\boldsymbol{X}}^2}\right)\right).$$

Thus choosing $\epsilon_1 \equiv 2Mc_{\boldsymbol{X}}^2 c_{n,m}$, where $c_{n,m} \equiv \max\left\{\frac{2(\log m^\tau + \log 6)}{n}, \sqrt{\frac{2(\log m^\tau + \log 6)}{n}}\right\}$, for $\gamma_{kj} \leq \epsilon_1/2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) + \gamma_{kj}\mathbb{1}_{k=\ell} - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| > \epsilon_1\right) \\
\leq &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| > \epsilon_1/2\right) \qquad (32) \\
\leq &2\exp\left(-\min\left(\frac{n\epsilon_1^2}{8M^2 c_{\boldsymbol{X}}^4}, \frac{n\epsilon_1}{4Mc_{\boldsymbol{X}}^2}\right)\right) \leq \frac{1}{3m^\tau}. \qquad (33)
\end{aligned}$$

Denote the event inside the probability in (32) as $\mathcal{E}_{k,\ell,j}(\epsilon_1/2)$.
By definition,

$$c_{\boldsymbol{X}}^2 = 4\max_k(4\Sigma_{kk} + 4\sqrt{e}\sqrt{\Sigma_{kk}}\mathbb{E}_0 X_k + e(\mathbb{E}_0 X_k)^2) \geq 4e\max_k(\Sigma_{kk} + (\mathbb{E}_0 X_k)^2).$$

By Lemmas 21.2 and 22.1, $\mathrm{var}(X_k) \leq \Sigma_{kk}$, so $c_{\boldsymbol{X}}^2 \geq 4e\max_k \mathbb{E}_0 X_k^2 \geq 4e\mathbb{E}_0 X_k^2 h_j(X_j)/M$. Thus, by $\epsilon_1 = 2Mc_{\boldsymbol{X}}^2 c_{n,m}$, on $\mathcal{E}_{k,k,j}(\epsilon_1/2)^\complement$ we have

$$\frac{1}{n}\sum_{i=1}^n {X_k^{(i)}}^2 h_j(X_j^{(i)}) \leq \mathbb{E}_0 X_k^2 h_j(X_j) + \epsilon_1/2 \leq \frac{\epsilon_1}{2}\left(1 + \frac{1}{4ec_{n,m}}\right).$$

Then

$$\frac{1}{1 + 1/(4ec_{n,m})}\frac{1}{n}\sum_{i=1}^n {X_k^{(i)}}^2 h_j(X_j^{(i)}) \leq \epsilon_1/2 \qquad (34)$$

on $\mathcal{E}_{k,k,j}(\epsilon_1/2)^\complement$, where we recall that $c_{n,m} \equiv \max\left\{\frac{2(\log m^\tau + \log 6)}{n}, \sqrt{\frac{2(\log m^\tau + \log 6)}{n}}\right\}$. Note that the multiplier on the left of (34) is increasing in $c_{n,m}$, and that $\frac{2(\log m^\tau + \log 6)}{n} > \frac{6\log m}{n}$ by the assumption that $\tau > 3$. Thus if we let

$$\gamma_{kj} \equiv \frac{1}{1 + 1/\left(4e\max\left\{6\log m/n, \sqrt{6\log m/n}\right\}\right)}\frac{1}{n}\sum_{i=1}^n {X_k^{(i)}}^2 h_j(X_j^{(i)}),$$

which is just a constant multiple of the (k,k)-th entry of $\boldsymbol{\Gamma}_j$ itself, with the constant explicitly calculable and depend on p and n only, then for $k = \ell$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) + \gamma_{kj} - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| \geq \epsilon_1\right) \leq \mathbb{P}(\mathcal{E}_{k,\ell,j}(\epsilon_1/2)) \leq \frac{1}{3m^\tau}.$$

Since this also holds for $k \neq \ell$ without the γ_{kj} term, by a union bound over m^3 events,

$$\mathbb{P}\left(\max_{j,k,\ell}\left|\frac{1}{n}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}) + \gamma_{kj}\mathbb{1}_{k=\ell} - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| \geq \epsilon_1\right) \leq \frac{1}{3m^{\tau-3}}. \qquad (35)$$

Now on the other hand, Lemma 22.1 and Hoeffding's inequality give for any $t_{2,1}, t_{2,2} > 0$ that,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j'(X_j^{(i)}) - \mathbb{E}_0 X_k h_j'(X_j)\right| \geq t_{2,1}\right) \leq 2\exp\left(-\frac{nt_{2,1}^2}{2M'^2 c_{\boldsymbol{X}}^2}\right),$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}) - \mathbb{E}_0 h_j(X_j)\right| \geq t_{2,2}\right) \leq 2\exp\left(-2nt_{2,2}^2/M^2\right).$$

Choosing $\epsilon_{2,1} \equiv \sqrt{2}M' c_{\boldsymbol{X}} \sqrt{\frac{\log m^{\tau-1} + \log 6}{n}}$, $\epsilon_{2,2} \equiv M\sqrt{\frac{\log m^{\tau-2} + \log 6}{2n}}$ and taking union bounds over m^2, and m events, respectively, we have

$$\mathbb{P}\left(\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j'(X_j^{(i)}) - \mathbb{E}_0 X_k h_j'(X_j)\right| \geq \epsilon_{2,1}\right) \leq \frac{1}{3m^{\tau-3}}, \tag{36}$$
$$\mathbb{P}\left(\max_{j}\left|\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}) - \mathbb{E}_0 h_j(X_j)\right| \geq \epsilon_{2,2}\right) \leq \frac{1}{3m^{\tau-3}}. \tag{37}$$

Hence, by (35) (36) (37), with probability at least $1 - m^{3-\tau}$, $\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty < \epsilon_1$ and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2 \equiv \epsilon_{2,1} + \epsilon_{2,2}$. Consider any $\tau > 3$, and let

$$c_2 \equiv \frac{6}{\alpha} c_{\boldsymbol{\Gamma}_0}, \qquad n \geq \max\{2M^2 c_{\boldsymbol{X}}^4 c_2^2 d_{\mathbf{K}_0}^2 (\tau \log m + \log 6), 2M c_{\boldsymbol{X}}^2 c_2 d_{\mathbf{K}_0} (\tau \log m + \log 6)\},$$
$$\begin{aligned}\lambda &> \frac{3(2-\alpha)}{\alpha} \max\{c_{\mathbf{K}_0}\epsilon_1, \epsilon_2\} \\ &\equiv \frac{3(2-\alpha)}{\alpha} \max\left\{4M c_{\mathbf{K}_0} c_{\boldsymbol{X}}^2 \frac{(\log m^\tau + \log 6)}{n}, \right. \\ &\qquad \left. 2M c_{\mathbf{K}_0} c_{\boldsymbol{X}}^2 \sqrt{\frac{2(\log m^\tau + \log 6)}{n}}, \sqrt{2}M' c_{\boldsymbol{X}} \sqrt{\frac{\log m^{\tau-1} + \log 6}{n}} + M\sqrt{\frac{\log m^{\tau-2} + \log 6}{2n}}\right\}.\end{aligned}$$

Then $d_{\mathbf{K}_0}\epsilon_1 \leq \alpha/(6c_{\boldsymbol{\Gamma}_0})$ and the results follow from Theorem 12. ■

Proof [Proof of Theorem 16] Similar to the proof of Theorem 15, by Theorem 12 it suffices to prove that for any $\tau > 3$, we can bound $\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty$ by some ϵ_1 and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty$ by some ϵ_2, uniformly with probability $1 - m^{3-\tau}$. Recall that $\boldsymbol{\Gamma} \in \mathbb{R}^{(m^2+m)\times(m^2+m)}$ is a rearrangement of $\boldsymbol{\Gamma}^{(*)}$, which is in turn formed by $\boldsymbol{\Gamma}_{11} \in \mathbb{R}^{m^2\times m^2}$, $\boldsymbol{\Gamma}_{12} \in \mathbb{R}^{m^2\times m}$, $\boldsymbol{\Gamma}_{12}^\top$ and $\boldsymbol{\Gamma}_{22} \in \mathbb{R}^{m\times m}$, all of which are block-diagonal with m blocks.
The j^{th} block of $\boldsymbol{\Gamma}_{11} \in \mathbb{R}^{m^2\times m^2}$ has (k,ℓ)-th entry

$$\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j(X_j^{(i)}),$$

the k^{th} entry in the j^{th} block of $\boldsymbol{\Gamma}_{12}$ is

$$-\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j(X_j^{(i)}),$$

the j^{th} diagonal entry of $\mathbf{\Gamma}_{22}$ is

$$\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}).$$

On the other hand $\boldsymbol{g} \in \mathbb{R}^{(m^2+m)}$ is a rearrangement of $\boldsymbol{g}^{(*)} \equiv [\boldsymbol{g}_1^\top, \boldsymbol{g}_2^\top]^\top$, where the entry in $\boldsymbol{g}_1 \in \mathbb{R}^{m^2}$ (obtained by linearizing a $m \times m$ matrix) corresponding to (j,k), is

$$\frac{1}{n}\sum_{i=1}^{n} X_k^{(i)} h_j'(X_j^{(i)}) + \mathbb{1}_{j=k}\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^{(i)}),$$

while the j-th component of $\boldsymbol{g}_2 \in \mathbb{R}^m$ is

$$-\frac{1}{n}\sum_{i=1}^{n} h_j'(X_j^{(i)}).$$

Recalling that the bounds in Lemma 22 also hold when $\boldsymbol{\mu} \neq 0$, we may then use bounds similar to those in the proof of Theorem 15, and use union bounds to arrive at analogous consistency results, modulus different constants. The amplifiers $\boldsymbol{\gamma}$ can be incorporated analogously. ■

Acknowledgments

This work was partially supported by grant DMS 1561814 from the US National Science Foundation.

Appendix A. Auxillary Lemmas and Definitions

In this appendix, to simplify notation, when it is clear from the context, the operator $\mathbb{E}$ is defined as the expectation under the true distribution, unless otherwise noted.
The following lemma verifies (A1)-(A2) in the case of truncated Gaussian distributions.

Lemma 19 (Assumptions for truncated Gaussian) *Consider the non-centered truncated Gaussian distribution defined by Equation (23) with unknown positive definite inverse covariance parameter $\mathbf{K}_0$ and unknown mean parameter $\boldsymbol{\mu}_0$. Then assuming $0 \leq h_j \leq M_j$, $\lim_{x_j \searrow 0^+} h_j(x_j) = 0$ and $|h_j'| \leq M_j'$, assumptions (A1)-(A2) for score matching are satisfied for any proposed parameters $\mathbf{K} \succ \mathbf{0}$ and $\boldsymbol{\mu}$. Taking $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_0 \equiv \mathbf{0}$ the assumptions also hold in the centered setting. Choosing $m = 1$ gives the univariate case.*

Proof [Proof of Lemma 19] Consider $p \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{K})$, with $\boldsymbol{k}_j$ the j-th column of $\mathbf{K}$. Let $M \equiv \max_j M_j$ and $M' \equiv \max_j M_j'$.

(A1) For any fixed $\boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}$ and any $p \in \mathcal{P}_+$ with parameters $\mathbf{K}$ and $\boldsymbol{\mu}$,

$$\lim_{x_j \nearrow \infty} h_j(x_j) p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \propto \lim_{x_j \nearrow \infty} h_j(x_j) \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^\top \mathbf{K}_0 (\boldsymbol{x}-\boldsymbol{\mu}_0)\right) \boldsymbol{k}_j^\top (\boldsymbol{x}-\boldsymbol{\mu})$$

$$= \lim_{x_j \nearrow \infty} h_j(x_j) \exp\left(C_1 + C_2 x_j - \frac{1}{2}\kappa_{0,jj} x_j^2\right)(C_3 + C_4 x_j)$$

for some constants C_1, C_2, C_3, and C_4 depending on $\boldsymbol{x}_{-j}$, $\mathbf{K}_0$, $\mathbf{K}$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}$. Since $\kappa_{0,jj} > 0$ and we assumed h_j to be bounded, the limit equals to 0 for all j and $\boldsymbol{x}_{-j}$. Similarly,

$$\begin{aligned}\lim_{x_j \searrow 0^+} h_j(x_j) p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) &\propto \lim_{x_j \searrow 0^+} h_j(x_j) \exp\left(C_1 + C_2 x_j - \frac{1}{2}\kappa_{0,jj} x_j^2\right)(C_3 + C_4 x_j) \\ &= \exp(C_1) C_3 \lim_{x_j \searrow 0^+} h_j(x_j) = 0\end{aligned}$$

if and only if we assume $\lim_{x_j \searrow 0^+} h_j(x_j) = 0$.

(A2) For any $p \in \mathcal{P}_+$ with parameters $\mathbf{K}$ and $\boldsymbol{\mu}$,

$$\begin{aligned}&\mathbb{E}_{p_0} \|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 \\ &\quad \le M \mathbb{E}_{p_0} \|\nabla \log p(\boldsymbol{X})\|_2^2 \;=\; M \mathrm{tr}\left(\mathbb{E}_{p_0}\left[(\mathbf{K}(\boldsymbol{X} - \boldsymbol{\mu}))(\mathbf{K}(\boldsymbol{X} - \boldsymbol{\mu}))^\top\right]\right) \\ &\quad = M \mathrm{tr}\left(\mathbf{K} \mathbb{E}_{p_0}\left[(\boldsymbol{X} - \boldsymbol{\mu}_0 + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}))(\boldsymbol{X} - \boldsymbol{\mu}_0 + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}))^\top\right] \mathbf{K}^\top\right) \\ &\quad = M \mathrm{tr}\left(\mathbf{K}\left(\mathbf{K}_0^{-1} + (\boldsymbol{\mu}_0 - \boldsymbol{\mu})(\boldsymbol{\mu}_0 - \boldsymbol{\mu})^\top\right)\mathbf{K}\right) < +\infty\end{aligned}$$

since M, $\mathbf{K}$, $\mathbf{K}_0$, $\boldsymbol{\mu}$, $\boldsymbol{\mu}_0$ are all finite constants. We also have

$$\begin{aligned}\mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 &= \sum_{i=1}^m \mathbb{E}_{p_0} \left|h_j'(X_j)\partial_j \log p(\boldsymbol{X}) + h_j(X_j)\partial_j^2 \log p(\boldsymbol{X})\right| \\ &\le \sum_{i=1}^m \mathbb{E}_{p_0} |h_j'(X_j)\partial_j \log p(\boldsymbol{X})| + \mathbb{E}_{p_0} |h_j(X_j)\partial_j^2 \log p(\boldsymbol{X})| \\ &\le \sum_{i=1}^m M' \mathbb{E}_{p_0} |\boldsymbol{k}_j^\top (\boldsymbol{X} - \boldsymbol{\mu})| + M\kappa_{jj} \\ &\le \sum_{i=1}^m M' |\boldsymbol{k}_j|^\top \mathbb{E}_{p_0} \boldsymbol{X} + M' |\boldsymbol{k}_j^\top \boldsymbol{\mu}| + M \mathrm{tr}(\mathbf{K}) < +\infty.\end{aligned}$$

Hence, (A1) and (A2) are both satisfied. ■

Definition 20 (Sub-Gaussian and Sub-Exponential Variables) *The* sub-gaussian ($r = 2$) *and* sub-exponential ($r = 1$) norms *of a random variable are defined as*

$$\|X\|_{\psi_r} \equiv \sup_{q \ge 1} q^{-1/r} (\mathbb{E}|X|^{rq})^{1/(rq)} \equiv \sup_{q \ge 1} q^{-1/r} \|X\|_{rq}.$$

If $\|X\|_{\psi_2} < \infty$ *we say* X *is* sub-gaussian*; if* $\|X\|_{\psi_1} < \infty$ *we call* X sub-exponential.
For a zero-mean *sub-gaussian random variable* X *also define the* sub-gaussian parameter

$$\tau(X) = \inf\{\tau \ge 0 : \mathbb{E} \exp(tX) \le \exp(\tau^2 t^2/2),\ \forall t \in \mathbb{R}\}.$$

Note that the definition of sub-gaussian norm given here allows the variable to be non-centered, and is different from the one in Vershynin (2010), which uses $\|X\|_q$ in the definition. Instead, it coincides with θ_2 in Buldygin and Kozachenko (2000). The definition of the sub-gaussian parameter is the same as in Buldygin and Kozachenko (2000), and the definition of the sub-exponential norm is as in Vershynin (2010).

Lemma 21 (Properties of Sub-Gaussian and Sub-Exponential Variables)

1) For any X and $r = 1, 2$, $\|X - \mathbb{E}X\|_{\psi_r} \leq 2\|X\|_{\psi_r}$ and $\|X\|_{\psi_r} \leq \|X - \mathbb{E}X\|_{\psi_r} + |\mathbb{E}X|$, as long as the expectation and norms are finite.

2) (Buldygin and Kozachenko, 2000) $\tau(X)$ is a norm on the space of all zero-mean sub-gaussian variables; in particular, $\tau(X + Y) \leq \tau(X) + \tau(Y)$ as long as the quantities are defined and finite.
If X is zero-mean sub-gaussian, then $\mathrm{var}(X) \leq \tau^2(X)$, $\|X\|_{\psi_2} \leq 2\tau(X)/\sqrt{e}$, $\tau(X) \leq \sqrt{e}\|X\|_{\psi_2}$.
If $X_1, \ldots, X_n$ are i.i.d. zero-mean sub-gaussian, $\tau\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \leq \frac{1}{\sqrt{n}}\tau(X_i)$.

3) If random variables X_1 and X_2 (not necessarily independent) are sub-gaussian with $\|X_1\|_{\psi_2} \leq K_1$ and $\|X_2\|_{\psi_2} \leq K_2$, then X_1X_2 is sub-exponential with $\|X_1X_2\|_{\psi_1} \leq K_1K_2$.

4) (Buldygin and Kozachenko, 2000) If X is zero-mean sub-gaussian, then

$$\mathbb{E}|X|^q \leq 2(q/e)^{q/2}\tau^q(X)$$

for any $q > 0$.

5) (Buldygin and Kozachenko, 2000) If $X_1, \ldots, X_n$ are independent zero-mean sub-gaussian variables, then for any $\epsilon > 0$,

$$\mathbb{P}(|X_1| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\tau^2(X_1)}\right),$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\max_i \tau^2(X_i)}\right).$$

6) (Vershynin, 2010) If $X_1, \ldots, X_n$ are independent zero-mean sub-exponential random variables with $K \geq \max_i \|X_i\|_{\psi_1}$, then for any $\epsilon > 0$,

$$\mathbb{P}(|X_1| \geq \epsilon) \leq 2\exp\left(-\min\left(\frac{\epsilon^2}{8e^2K^2}, \frac{\epsilon}{4eK}\right)\right),$$
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{8e^2K^2}, \frac{n\epsilon}{4eK}\right)\right).$$

7) (Boucheron et al., 2013) If for X_i i.i.d. there exists some $B > 0$ such that

$$\sup_{q \geq 2}\left(\frac{\mathbb{E}|X|^q}{q!}\right)^{1/q} \leq B/2$$

then for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{2B^2}, \frac{n\epsilon}{2B}\right)\right).$$

Proof

1) For $r = 1, 2$, by triangle inequality, $\|X - \mathbb{E}X\|_{\psi_r} \leq \|X\|_{\psi_r} + \|\mathbb{E}X\|_{\psi_r} = \|X\|_{\psi_r} + |\mathbb{E}X| \leq \|X\|_{\psi_r} + \mathbb{E}|X| \leq 2\|X\|_{\psi_r}$, where in the last step we used the definition of $\|\cdot\|_{\psi_r}$ with $q = 1$ for $r = 1$ and $\mathbb{E}|X| \leq (\mathbb{E}|X|^2)^{1/2}$ with $q = 2$ for $r = 2$. On the other hand, $\|X\|_{\psi_r} \leq \|X - \mathbb{E}X\|_{\psi_r} + \|\mathbb{E}X\|_{\psi_r} = \|X - \mathbb{E}X\|_{\psi_r} + |\mathbb{E}X|$.

2) These follow from Theorems 1.2 and 1.3 and Lemmas 1.2 and 1.7 from Buldygin and Kozachenko (2000), and $\sqrt[4]{3.1}e^{9/16}/\sqrt{2} \approx 1.6467 \leq 1.6487 \approx \sqrt{e}$.

3) By Hölder's inequality (or Cauchy-Schwarz),

$$\begin{aligned}
\|X_1X_2\|_{\psi_1} &= \sup_{q\geq 1} q^{-1}(\mathbb{E}|X_1X_2|^q)^{1/q} = \sup_{q\geq 1} q^{-1}(\mathbb{E}|X_1^qX_2^q|)^{1/q} \\
&\leq \sup_{q\geq 1} q^{-1}\left[(\mathbb{E}|X_1|^{2q})^{1/2}(\mathbb{E}|X_2|^{2q})^{1/2}\right]^{1/q} \\
&\leq \sup_{q\geq 1}\left[q^{-1/2}(\mathbb{E}|X_1|^{2q})^{1/2q}\right]\sup_{q\geq 1}\left[q^{-1/2}(\mathbb{E}|X_2|^{2q})^{1/2q}\right] \\
&= \|X_1\|_{\psi_2}\|X_2\|_{\psi_2} \leq K_1K_2.
\end{aligned}$$

4) This is Lemma 1.4 from Buldygin and Kozachenko (2000).

5) This is Theorem 1.5 from Buldygin and Kozachenko (2000).

6) This follows from Corollary 5.17 from Vershynin (2010).

7) By Theorem 2.10 of Boucheron et al. (2013) wherein we let $v \equiv nB^2/2$ and $c \equiv B/2$, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{B^2 + B\epsilon}\right)$$

for all $\epsilon > 0$. (Theorem 2.10 gives an one-sided bound; bound for the other side is obtained by taking $X_i = -X_i$). The inequality follows by splitting into cases $\epsilon \leq B$ and $\epsilon > B$.

■

Lemma 22 *Suppose* $\boldsymbol{X}$ *follows a truncated normal distribution on* $\mathbb{R}_+^m$ *with parameters* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma} = \mathbf{K}^{-1} \succ \mathbf{0}$. *Let* $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ *be i.i.d. copies of* $\boldsymbol{X}$, *with* j*-th component of the* i*-th copy being* $X_j^{(i)}$. *Then*

1) For $j = 1, \ldots, p$, $\tau(X_j - \mathbb{E}X_j) \leq \sqrt{\Sigma_{jj}}$. That is, the sub-gaussian parameter of any marginal distribution of $\boldsymbol{X}$, after centering, is bounded by the square root of its corresponding diagonal entry in the covariance parameter $\boldsymbol{\Sigma}$. Then for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} - \mathbb{E}X_j\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\Sigma_{jj}}\right).$$

In particular, if h_0 is a function bounded by M_0, then for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8M_0^2(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j)^2}\right),$$

$$\tau\left(\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right) \leq \frac{2M_0}{\sqrt{n}}(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j),$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right\|_{\psi_2} \leq \frac{4M_0}{\sqrt{en}}(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j).$$

2) For $j, k, \ell \in \{1, \ldots, p\}$, if h_0 is a function bounded by M_0, then

$$\|X_j X_k h_0(X_\ell) - \mathbb{E}X_j X_k h_0(X_\ell)\|_{\psi_1} \leq \frac{M_0}{2e} c_{\boldsymbol{X}}^2 \tag{38}$$

with $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j)$. In particular, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} h_0(X_\ell^{(i)}) - \mathbb{E}X_j X_k h_0(X_\ell)\right| > \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{2M_0^2 c_{\boldsymbol{X}}^4}, \frac{n\epsilon}{2M_0 c_{\boldsymbol{X}}^2}\right)\right).$$

Proof [Proof of Lemma 22]

1) Without loss of generality choose $j = 1$. By the definition of sub-gaussian parameters, we need to show that for all $t \in \mathbb{R}$,

$$\mathbb{E}\exp(tX_1) \leq \exp(t^2\Sigma_{11}/2 + t\mathbb{E}X_1),$$

which is equivalent to

$$t^2\Sigma_{11}/2 + t\mathbb{E}X_1 - \log\mathbb{E}\exp(tX_1) \geq 0 \quad \forall t \in \mathbb{R}. \tag{39}$$

Since the left-hand side of (39) equals 0 at $t = 0$, it suffices to show that its derivative

$$t\Sigma_{11} + \mathbb{E}X_1 - \frac{\mathrm{d}\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t} = t\Sigma_{11} + \mathbb{E}X_1 - \frac{\frac{\mathrm{d}\mathbb{E}\exp(tX_1)}{\mathrm{d}t}}{\mathbb{E}\exp(tX_1)} \tag{40}$$

is non-negative on $(0, \infty)$ and non-positive on $(-\infty, 0)$. By properties of moment-generating functions, $\frac{\mathrm{d}\mathbb{E}\exp(tX_1)}{\mathrm{d}t}$ evaluated at $t = 0$ equals $\mathbb{E}X_1$, so (40) equals 0 at $t = 0$. It in turn suffices to show the derivative of (40), namely

$$\Sigma_{11} - \frac{\mathrm{d}^2\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t^2} \tag{41}$$

is non-negative in $t \in \mathbb{R}$.

By Tallis (1961), denoting the first column of $\mathbf{\Sigma}$ as $\mathbf{\Sigma}_1$, the moment-generating function of the marginal distribution of X_1 is

$$\frac{\int_{\mathbb{R}_+^p - \boldsymbol{\mu} - t\mathbf{\Sigma}_1} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}_+^p - \boldsymbol{\mu}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x}} \exp\left(t\mu_1 + \frac{1}{2}t^2\Sigma_{11}^2\right).$$

The quantity in (41) thus becomes

$$-\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log \int_{\mathbb{R}_+^p - \boldsymbol{\mu} - t\mathbf{\Sigma}_1} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x}.$$

Showing this is non-negative in $t \in \mathbb{R}$ is equivalent to showing that the integral itself is log-concave in t. But

$$\int_{\mathbb{R}_+^p - \boldsymbol{\mu} - t\mathbf{\Sigma}_1} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{x}\right) \mathbb{1}_{\mathbb{R}_+^p - \boldsymbol{\mu}}(\boldsymbol{x} + t\mathbf{\Sigma}_1) \, \mathrm{d}\boldsymbol{x}$$

with $\exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{\Sigma}^{-1}\boldsymbol{x}\right)$ log-concave in $\boldsymbol{x}$ and $\mathbb{1}_{\mathbb{R}_+^p - \boldsymbol{\mu}}(\boldsymbol{x} + t\mathbf{\Sigma}_1)$ log-concave in $(\boldsymbol{x}, t)$ since $\mathbb{R}_+^p - \boldsymbol{\mu}$ is a convex set (half-space). Since log-concavity is closed under multiplication and integration over $\mathbb{R}^p$, the integral is indeed log-concave, and our proof of the bound on the sub-gaussian parameter of $X_j - \mathbb{E}X_j$ is complete. The tail bound follows from 5) of Lemma 21.

Now by 1) and 2) of Lemma 21,

$$\|X_j\|_{\psi_2} \leq 2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j.$$

If h_0 is a function bounded by M_0, then by definition

$$\|X_j h_0(X_k)\|_{\psi_2} \leq M_0 \left(2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j\right).$$

By 1) and 2) of Lemma 21 again,

$$\begin{aligned}
\tau(X_j h_0(X_k) - \mathbb{E}X_j h_0(X_k)) &\leq \sqrt{e}\|X_j h_0(X_k) - \mathbb{E}X_j h_0(X_k)\|_{\psi_2} \\
&\leq 2\sqrt{e}\|X_j h_0(X_k)\|_{\psi_2} \\
&\leq 2M_0(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j).
\end{aligned}$$

The tail bound thus follows from the first inequality using 5) of Lemma 21. By 2),

$$\begin{aligned}
\tau\left(\frac{1}{n}\sum_{i=1}^n X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right) &\leq \frac{2M_0}{\sqrt{n}}(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j), \\
\left\|\frac{1}{n}\sum_{i=1}^n X_j^{(i)} h_0(X_k^{(i)}) - \mathbb{E}X_j h_0(X_k)\right\|_{\psi_2} &\leq \frac{4M_0}{\sqrt{en}}(2\sqrt{\Sigma_{jj}} + \sqrt{e}\mathbb{E}X_j).
\end{aligned}$$

2) By the proof of 1) of this lemma, $\|X_j\|_{\psi_2} \le 2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j$, and by 3) of Lemma 21,

$$\|X_j X_k\|_{\psi_1} \le (2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j)(2\sqrt{\Sigma_{kk}/e} + \mathbb{E}X_k) \le \max_j (2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j)^2.$$

Since h_0 is a function bounded by M_0, by definition

$$\|X_j X_k h_0(X_\ell)\|_{\psi_1} \le M_0 \max_j (2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j)^2.$$

Then by 1) of Lemma 21 again,

$$\|X_j X_k h_0(X_\ell) - \mathbb{E}X_j X_k h_0(X_\ell)\|_{\psi_1} \le 2M_0 \max_j (2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j)^2.$$

The tail bound then follows from 6) of Lemma 21.

■

Although not used in our analysis for the consistency results, in the special case of $h_0 \equiv 1$, we also have the following lemma. The notable difference between bounds (42) below and (38) from Lemma 22.2 is in the constants and dependency on $\mathbb{E}X_j$: The constants in the denominator in the right-hand side of (38) is smaller and thus gives a tighter bound, but (42) is preferred when $\mathbb{E}X_j$ is notably large compared to $\sqrt{\Sigma_{jj}}$, since the constant is only linear in $\mathbb{E}X_j$.

Lemma 23 *Consider the setting in Lemma 22. Then for $j, k \in \{1, \dots, p\}$, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n X_j^{(i)} X_k^{(i)} - \mathbb{E}X_j X_k\right| \ge \epsilon\right) \le 4\exp\left(-\min\left(\frac{2n\epsilon^2}{C_1^2}, \frac{n\epsilon}{C_1}\right)\right), \tag{42}$$

where $C_1 \equiv 91 \max_j \Sigma_{jj} + 72 \max_j \mathbb{E}X_j \max_j \sqrt{\Sigma_{jj}}$.

Proof [Proof of Lemma 23] We use a proof similar to Lemma 1 in Ravikumar et al. (2011) (note that the assumption that $\mathbb{E}X_j = 0$ in Ravikumar et al. (2011) does not hold in our case). Define

$$U_{jk}^{(i)} \equiv X_j^{(i)} + X_k^{(i)}, \quad U_{jk} \equiv X_j + X_k, \quad V_{jk}^{(i)} \equiv X_j^{(i)} - X_k^{(i)}, \quad V_{jk} \equiv X_j - X_k.$$

Since $X_j^{(i)} X_k^{(i)} = \frac{1}{4}\left({U_{jk}^{(i)}}^2 - {V_{jk}^{(i)}}^2\right)$, by union bound we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_j^{(i)} X_k^{(i)} - \mathbb{E}X_j X_k\right| \ge \epsilon\right) \\
\le &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n {U_{jk}^{(i)}}^2 - \mathbb{E}U_{jk}^2\right| \ge 2\epsilon\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n {V_{jk}^{(i)}}^2 - \mathbb{E}V_{jk}^2\right| \ge 2\epsilon\right).
\end{aligned}$$

We next define

$$Z_{jk}^{(i)} \equiv {U_{jk}^{(i)}}^2 - \mathbb{E}U_{jk}^2 = {A_{jk}^{(i)}}^2 + B_{jk}^{(i)} + C_{jk}, \qquad \overline{X}_j^{(i)} \equiv X_j^{(i)} - \mathbb{E}X_j,$$

$$A_{jk}^{(i)} \equiv \overline{X}_j^{(i)} + \overline{X}_k^{(i)}, \quad B_{jk}^{(i)} \equiv 2(\mathbb{E}X_j + \mathbb{E}X_k)(\overline{X}_j^{(i)} + \overline{X}_k^{(i)}), \quad C_{jk} \equiv -\mathbb{E}(\overline{X}_j^{(i)} + \overline{X}_k^{(i)})^2.$$

Then since τ is a norm by 2) of Lemma 21, A_{jk} is sub-gaussian with parameter $\leq \sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}$, and B_{jk} is sub-gaussian with parameter $\leq 2(\mathbb{E}X_j + \mathbb{E}X_k)(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}})$. Using 4) of Lemma 21 together with the inequality $(a+b+c)^q \leq (3\max\{a,b,c\})^q \leq 3^q(a^q + b^q + c^q)$ for all $a, b, c \geq 0$ and $q > 0$, we have for any $q \geq 2$

$$\begin{aligned}
(\mathbb{E}|Z_{jk}|^q)^{1/q} &\leq \left(3^m \left(\mathbb{E}A_{jk}^{2q} + \mathbb{E}|B_{jk}|^q + |C_{jk}|^q\right)\right)^{1/q} \\
&\leq 3^{1+1/q}\left((\mathbb{E}|A_{jk}|^{2q})^{1/q} + (\mathbb{E}|B_{jk}|^q)^{1/q} + |C_{jk}|\right) \\
&\leq 3^{1+1/q}\Big(2^{1/q}(2q/e)(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}})^2 \\
&\qquad +2^{1/q}\sqrt{q/e}2(\mathbb{E}X_j + \mathbb{E}X_k)(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}) + \mathrm{var}(X_j + X_k)\Big).
\end{aligned}$$

Using $\mathrm{var}(X+Y) \leq 2(\mathrm{var}(X) + \mathrm{var}(Y))$ and the fact that

$$\mathrm{var}(X_j) = \mathrm{var}(X_j - \mathbb{E}X_j) \leq \tau^2(X_j - \mathbb{E}X_j) \leq \Sigma_{jj}$$

(by 2) of Lemma 21 and part 1 of this theorem), we then have

$$\left(\frac{\mathbb{E}|Z_{jk}|^q}{q!}\right)^{1/q} \leq$$
$$3^{1+1/q}\frac{2^{3+1/q}(q/e)\max_j \Sigma_{jj} + 2^{3+1/q}\sqrt{q/e}\max_j \mathbb{E}X_j \cdot \max_j \sqrt{\Sigma_{jj}} + 4\max_j \Sigma_{jj}}{(q!)^{1/q}}.$$

Since all three coefficients involving q are decreasing in $q \geq 2$, we have

$$\sup_{q\geq 2}\left(\frac{\mathbb{E}|Z_{jk}|^q}{q!}\right)^{1/q} \leq (48\sqrt{3}/e + 6\sqrt{6})\max_j \Sigma_{jj} + 24\sqrt{6/e}\max_j \mathbb{E}X_j \max_j \sqrt{\Sigma_{jj}}.$$

Thus by 7) of Lemma 21, letting $B \equiv (91\max_j \Sigma_{jj} + 72\max_j \mathbb{E}X_j \max_j \sqrt{\Sigma_{jj}})$, we have for all $\epsilon > 0$:

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n Z_{jk}^{(i)}\right| \geq 2\epsilon\right) \leq 2\exp\left(-\min\left(\frac{2n\epsilon^2}{B^2}, \frac{n\epsilon}{B}\right)\right).$$

A tail bound for the sample average of V_{jk}^2 can be similarly derived, and the result follows. ■

Appendix B. Simulation Results for ER Graphs

Here we show simulation results for Erdös-Rényi graphs (defined in Section 6.2) under the same settings as in Section 6. Interpretations are similar to the SUB graphs and are omitted.

B.1 Choice of h

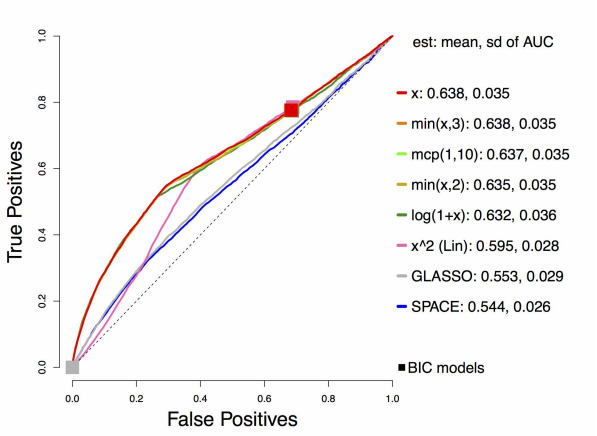
B.1.1 Truncated Centered GGMs

Centered, $n = 80$, multiplier 1.8647, ER					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.632	0.036	∞	0.638	0.035
2	0.632	0.036	3	0.638	0.035
1	0.630	0.035	2	0.635	0.035
0.5	0.613	0.033	1	0.623	0.033
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.637	0.035	10	0.638	0.035
5	0.636	0.036	5	0.637	0.035
1	0.617	0.033	2	0.632	0.035
$x^{1.5}$: (0.627, 0.032)			x^2: (0.595,0.028)		
GLASSO (0.553,0.029)			SPACE: (0.544, 0.026)		
NS: (0.543,0.028)			SJ: (0.519,0.028)		

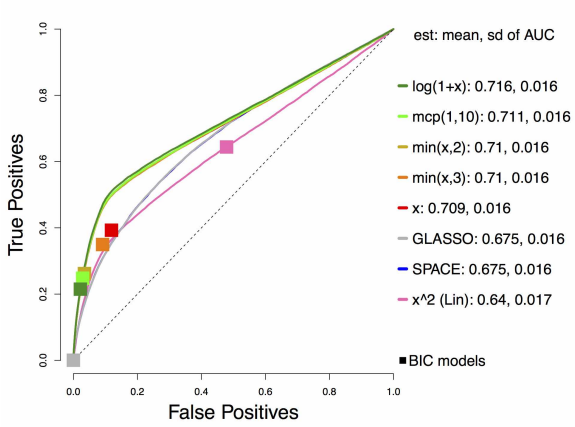
Centered, $n = 1000$, multiplier 1, ER					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.716	0.016	2	0.710	0.016
2	0.716	0.016	3	0.710	0.016
1	0.715	0.016	1	0.710	0.017
0.5	0.694	0.017	∞	0.709	0.016
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.714	0.016	2	0.713	0.016
10	0.711	0.016	5	0.711	0.016
1	0.707	0.017	10	0.710	0.016
$x^{1.5}$: (0.678,0.016)			x^2: (0.64,0.017)		
GLASSO: (0.675,0.016)			SPACE: (0.675,0.016)		
NS: (0.675,0.016)			SJ: (0.624,0.017)		

Centered, $n = 1000$, multiplier 1.6438, ER					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.796	0.014	∞	0.795	0.014
2	0.796	0.014	3	0.794	0.014
1	0.794	0.014	2	0.792	0.014
0.5	0.772	0.015	1	0.784	0.015
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.796	0.014	5	0.795	0.014
10	0.796	0.014	10	0.795	0.014
1	0.778	0.015	2	0.793	0.014
$x^{1.5}$: (0.757,0.015)			x^2: (0.693,0.016)		
GLASSO: (0.675,0.016)			SPACE: (0.675,0.016)		
NS: (0.675,0.016)			SJ: (0.624,0.017)		

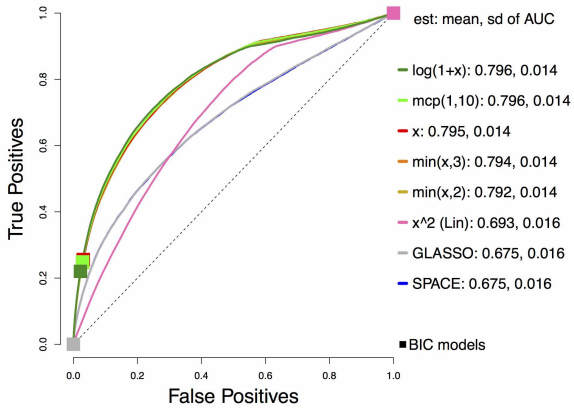
Table 4: Mean and standard deviation of areas under the ROC curves (AUC) using different estimators in the centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multipliers 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).



(a) $n = 80$, mult $= 1.8647$, ER



(b) $n = 1000$, mult $= 1$, ER



(c) $n = 1000$, mult $= 1.6438$, ER

Figure 10: Average ROC curves of our *centered* estimator for $m = 100$ variables and two sample sizes n under various choices of h, compared to SPACE and GLASSO, for the *truncated centered GGM* case. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.

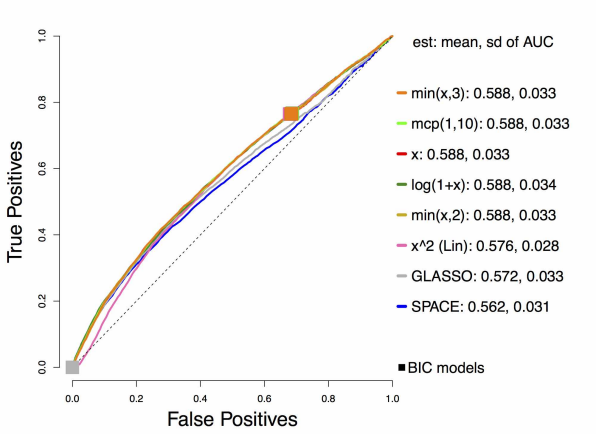
B.1.2 Truncated Non-Centered GGMs

Non-centered profiled, $n = 80$, multiplier 1.8647, ER					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
1	0.588	0.034	3	0.588	0.033
∞	0.588	0.034	∞	0.588	0.033
2	0.588	0.034	2	0.588	0.033
0.5	0.576	0.033	1	0.583	0.033
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.588	0.033	5	0.588	0.033
10	0.588	0.033	10	0.588	0.033
1	0.581	0.033	2	0.587	0.033
$x^{1.5}$: (0.582,0.028)			x^2: (0.576,0.028)		
GLASSO: (0.572,0.033)			SPACE: (0.562,0.031)		
NS: (0.560,0.032)			SJ: (0.535,0.027)		

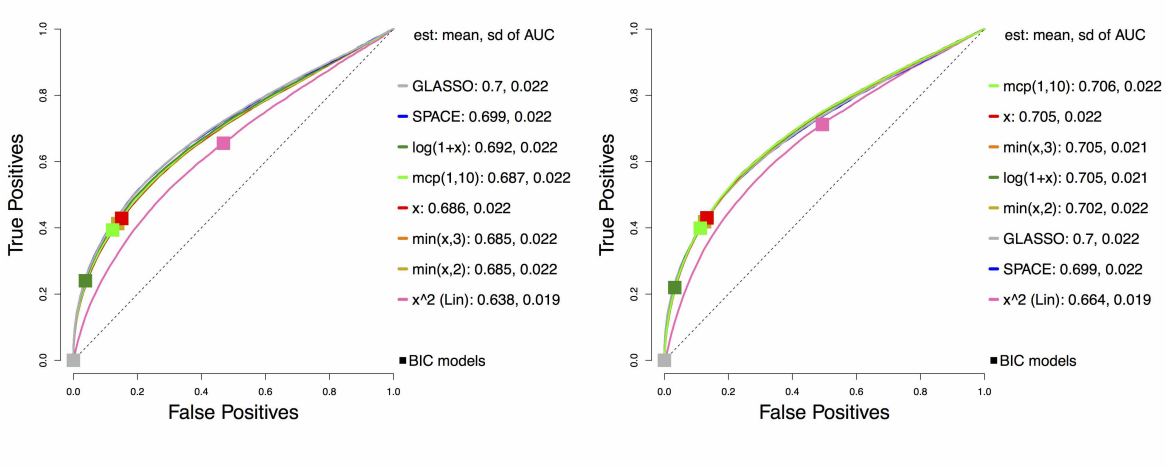
Non-centered profiled, $n = 1000$, multiplier 1, ER					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
2	0.692	0.022	1	0.687	0.022
∞	0.692	0.022	∞	0.686	0.022
1	0.691	0.022	3	0.685	0.022
0.5	0.684	0.02	2	0.685	0.022
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.689	0.022	2	0.687	0.022
1	0.689	0.020	5	0.687	0.022
10	0.687	0.022	10	0.686	0.022
$x^{1.5}$: (0.663,0.020)			x^2: (0.638,0.019)		
GLASSO (0.700,0.022)			SPACE: (0.699,0.022)		
NS: (0.699,0.022)			SJ: (0.655,0.021)		

Non-centered profiled, $n = 1000$, multiplier 1.6438, ER					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
2	0.705	0.021	∞	0.705	0.022
∞	0.705	0.021	3	0.705	0.021
1	0.703	0.021	2	0.702	0.022
0.5	0.683	0.019	1	0.695	0.021
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.706	0.021	10	0.705	0.022
10	0.706	0.022	5	0.705	0.022
1	0.690	0.019	2	0.703	0.022
$x^{1.5}$: (0.689,0.021)			x^2: (0.664,0.019)		
GLASSO (0.700,0.022)			SPACE: (0.699,0.022)		
NS: (0.699,0.022)			SJ: (0.655,0.021)		

Table 5: Mean and standard deviation of AUC using different *profiled* estimators in the non-centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multipliers 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).



(a) $n = 80$, mult $= 1.8647$, ER



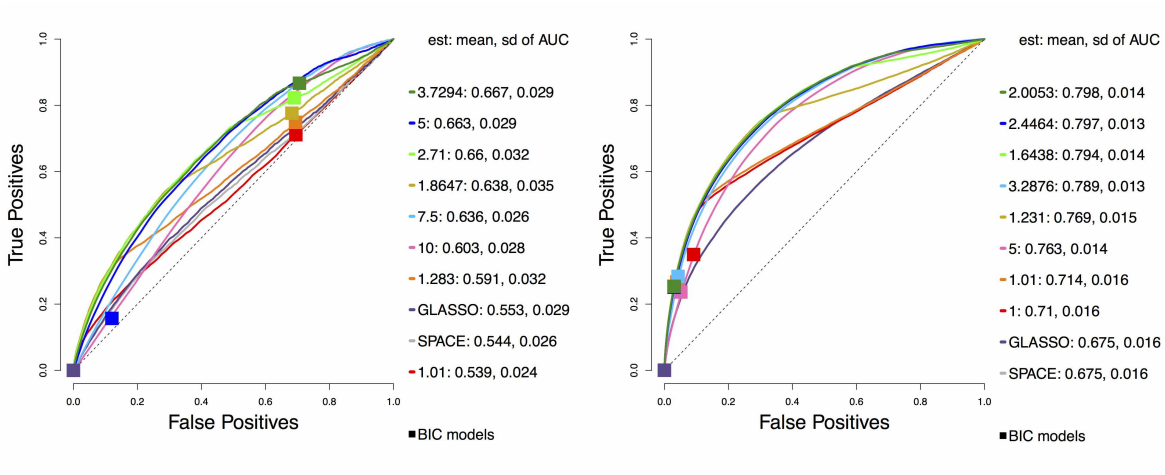
(b) $n = 1000$, mult $= 1$, ER

(c) $n = 1000$, mult $= 1.6438$, ER

Figure 11: Average ROC curves of our *non-centered profiled* estimator with various choices of h, compared to SPACE and GLASSO, for the *truncated non-centered GGM* case. $n = 80$ or 1000, $m = 100$. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.

B.2 Choice of multiplier

B.2.1 Truncated Centered GGMs



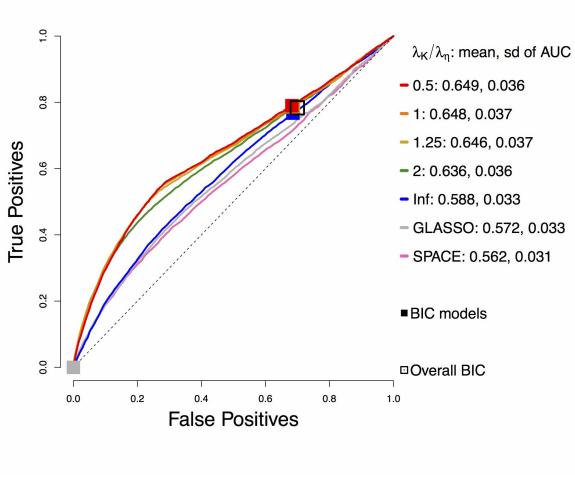
(a) $n = 80$, ER

(b) $n = 1000$, ER

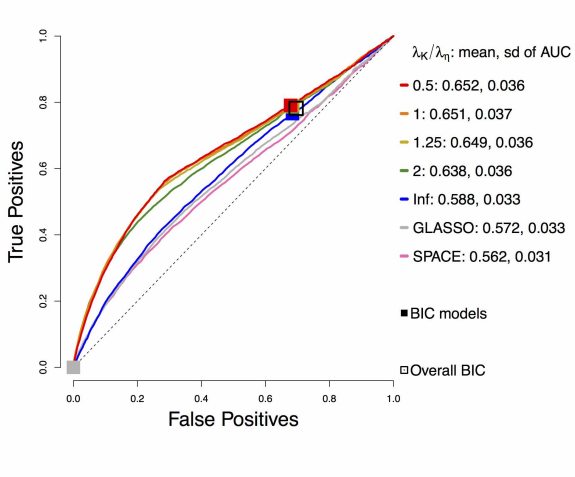
Figure 12: Performance of $\min(x, 3)$ for truncated centered GGMs using different multipliers, compared to GLASSO and SPACE, in the centered setting, $n = 80$ or 1000.

B.2.2 Truncated Non-Centered GGMs

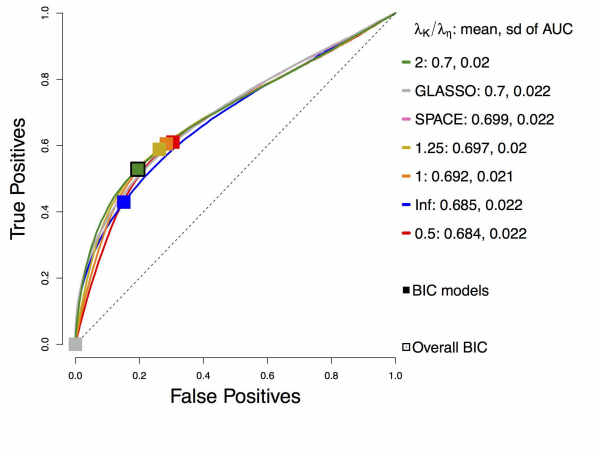
We recall that while for one run, the best model picked by BIC falls on the ROC curve, in (e) the red square is off the curve since it corresponds to the average of the true and false positive rates of the chosen BIC models over all 50 runs, and due to bimodality of the distribution of the chosen BIC models for different runs. But in all cases, the average of the models picked by BIC tuned over both $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$ looks reasonable.



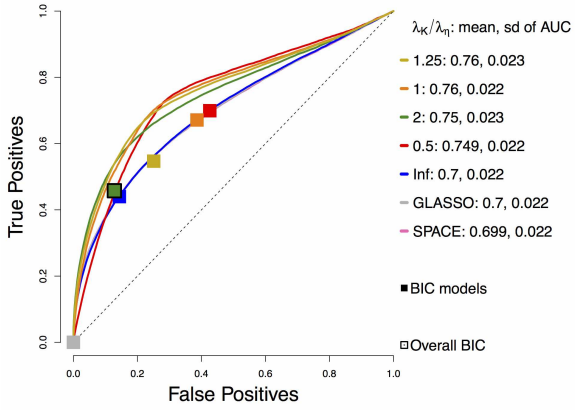
(a) $n = 80$, mult $= 1.7897$, ER



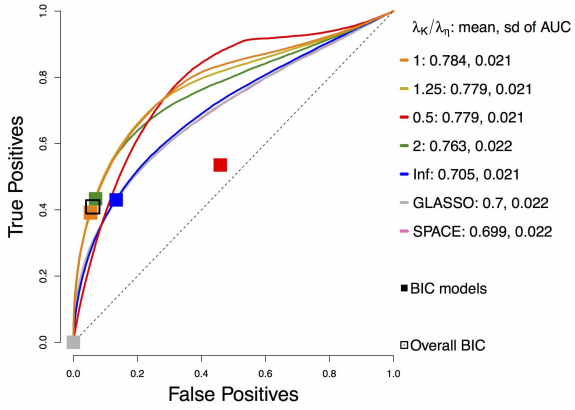
(b) $n = 80$, mult $= 1.8647$, ER



(c) $n = 1000$, mult $= 1$, ER



(d) $n = 1000$, mult $= 1.2310$, ER



(e) $n = 1000$, mult $= 1.6438$, ER

Figure 13: Performance of the non-centered estimator with $h(x) = \min(x, 3)$. Each curve corresponds to a different choice of $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$. Squares indicate models picked by eBIC with refit. The square with black outline has highest eBIC among all models (combinations of $\lambda_{\mathbf{K}}$, $\lambda\boldsymbol{\eta}$). The multipliers correspond to medium or high for $n = 80$, and low, medium and high for $n = 1000$, respectively.

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Igor Brikun, Deborah Nusskern, Daniel Gillen, Amy Lynn, Daniel Murtagh, John Feczko, William G Nelson, and Diha Freije. A panel of DNA methylation markers reveals extensive methylation in histologically benign prostate biopsy cores from cancer patients. *Biomarker Research*, 2(1):25, 2014.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Valeriui Vladimirovich Buldygin and IU V Kozachenko. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Soc., 2000.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Shizhe Chen, Daniela M Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2014.

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Mathias Drton and Michael D Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael Von Rhein, and Jan D Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):803–825, 2015.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854, 2016.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman, et al. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. 135:153–162, 2015.

Jun Luo, Thomas Dunn, Charles Ewing, Jurga Sauvageot, Yidong Chen, Jeffrey Trent, and William Isaacs. Gene expression signature of benign prostatic hyperplasia revealed by cDNA microarray analysis. *The Prostate*, 51(3):189–200, 2002.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Giuseppe Morgia, Mario Falsaperla, Grazia Malaponte, Massimo Madonia, Manuela Indelicato, Salvatore Travali, and Maria Clorinda Mazzarino. Matrix metalloproteinases as diagnostic (MMP-13) and prognostic (MMP-2, MMP-9) markers of prostate cancer. *Urological Research*, 33(1):44–50, 2005.

Shintaro Narita, Alan So, Susan Ettinger, Norihiro Hayashi, Mototsugu Muramaki, Ladan Fazli, Youngsoo Kim, and Martin E Gleave. GLI2 knockdown using an antisense oligonucleotide induces apoptosis and chemosensitizes cells to paclitaxel in androgen-independent prostate cancer. *Clinical Cancer Research*, 14(18):5769–5777, 2008.

Matthew Parry, A Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Mona Stearns and Mark E Stearns. Evidence for increased activated metalloproteinase 2 (mmp-2a) expression associated with human prostate cancer progression. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 8(2):69–75, 1996.

Georges M Tallis. The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 223–229, 1961.

Saravanan Thiyagarajan, Neehar Bhatia, Shannon Reagan-Shaw, Diana Cozma, Andrei Thomas-Tikhonenko, Nihal Ahmad, and Vladimir S Spiegelman. Role of GLI2 transcription factor in growth and tumorigenicity of prostate cells. *Cancer Research*, 67(22): 10642–10646, 2007.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2): 245–266, 2012.

Dominique Trudel, Yves Fradet, François Meyer, François Harel, and Bernard Têtu. Significance of MMP-2 expression in prostate cancer. *Cancer Research*, 63(23):8511–8515, 2003.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2013.

Susan R Wilson, Sandra Gallagher, Kate Warpeha, and Susan J Hawthorne. Amplification of MMP-2 and MMP-9 production by prostate cancer cell lines via activation of protease-activated receptors. *The Prostate*, 60(2):168–174, 2004.

Tiancheng Xie, Binbin Dong, Yangye Yan, Guanghui Hu, and Yunfei Xu. Association between MMP-2 expression and prostate cancer: A meta-analysis. *Biomedical Reports*, 4 (2):241–245, 2016.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

Ming Yu, Mladen Kolar, and Varun Gupta. Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, pages 2829–2837, 2016.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Graphical models for non-negative data using generalized score matching. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1790, 2018.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Teng Zhang and Hui Zou. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, 101(1):103–120, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.