Journal of Machine Learning Research 20 (2019) 1-70 Submitted 5/18; Revised 4/19; Published 4/19

Generalized Score Matching for Non-Negative Data

Shiqing Yu SQYU@UW.EDU
Department of Statistics
University of Washington, Seattle, WA, U.S.A.

Mathias Drton MD5@UW.EDU
Department of Mathematical Sciences
University of Copenhagen, Copenhagen, Denmark
and
Department of Statistics
University of Washington, Seattle, WA, U.S.A.

Ali Shojaie ASHOJAIE@UW.EDU
Department of Biostatistics
University of Washington, Seattle, WA, U.S.A.

Editor: Aapo Hyvarinen

Abstract

A common challenge in estimating parameters of probability density functions is the intractability of the normalizing constant. While in such cases maximum likelihood estimation may be implemented using numerical integration, the approach becomes computationally intensive. The score matching method of Hyvärinen (2005) avoids direct calculation of the normalizing constant and yields closed-form estimates for exponential families of continuous distributions over $\mathbb{R}^m$. 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. As an example, we consider a general class of pairwise interaction models. Addressing an overlooked inexistence problem, we generalize the regularized score matching method of Lin et al. (2016) and improve its theoretical guarantees for non-negative Gaussian graphical models.

Keywords: exponential family, graphical model, positive data, score matching, sparsity

1. Introduction

Score matching was first developed in Hyvärinen (2005) for continuous distributions supported on all of $\mathbb{R}^m$. Consider such a distribution P_0, with density p_0 and support equal to $\mathbb{R}^m$. Let $\mathcal{P}$ be a family of distributions with twice continuously 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 integration by parts 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. 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 integration by parts arguments underlying the score matching estimator may fail due to discontinuities at the boundary of the support. For data supported on the non-negative orthant $\mathbb{R}^m_+$, 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, both theoretically and empirically. To demonstrate these advantages, we consider a family of graphical models on $\mathbb{R}^m_+$, which does not have tractable normalizing constants and hence serves as a practical example.

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 entries 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., 2015; Yang et al., 2015), as well as semi- or non-parametric estimation techniques (Fellinghauer et al., 2013; Voorman et al., 2014). In this paper, we apply our method to a class of pairwise interaction models that generalizes non-negative Gaussian random variables, as recently considered by Lin et al. (2016) and Yu et al. (2016), as well as square root graphical models proposed by Inouye et al. (2016) when the sufficient statistic function is a pure power. However, our main ideas can also be applied for other classes of exponential families whose support is restricted to a rectangular set.

Our focus will be on *pairwise interaction power models* with probability distributions having (Lebesgue) densities proportional to

$$\exp\left\{-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a + \boldsymbol{\eta}^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b}\right\} \tag{1}$$

on $\mathbb{R}^m_+ \equiv [0,\infty)^m$. Here $a > 0$ and $b \geq 0$ are known constants, and $\mathbf{K} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\eta} \in \mathbb{R}^m$ are unknown parameters of interest. When $b = 0$ we define $(x^b - 1)/b \equiv \log x$ and $\mathbb{R}^m_+ \equiv (0,\infty)^m$. This class of models is motivated by the form of important univariate distributions for non-negative data, including gamma and truncated normal distributions. It provides a framework for pairwise interaction that is concrete yet rich enough to capture key differences in how densities may behave at the boundary of the non-negative orthant,

$\mathbb{R}_+^m$. Moreover, the conditional independence graph of a random vector $\boldsymbol{X}$ with distribution as in (1) 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 the interaction matrix $\mathbf{K}$. Section 5.1 gives further details on these models. We will develop estimators of $(\boldsymbol{\eta}, \mathbf{K})$ in (1) and the associated conditional independence graph using the proposed *generalized score matching*.

A special case of (1) are truncated Gaussian graphical models, with $a = b = 1$. Let $\boldsymbol{\mu} \in \mathbb{R}^m$, and let $\mathbf{K}$ be a positive definite matrix. Then a non-negative random vector $\boldsymbol{X}$ follows a truncated normal distribution for mean parameter $\boldsymbol{\mu}$ and inverse covariance parameter $\mathbf{K}$, in symbols $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{K})$, if it has density proportional to

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

on $\mathbb{R}_+^m$. We refer to $\boldsymbol{\Sigma} = \mathbf{K}^{-1}$ as the covariance parameter of the distribution, and note that the $\boldsymbol{\eta}$ parameter in (1) is $\mathbf{K}\boldsymbol{\mu}$. Another special case of (1) is the exponential square root graphical models in Inouye et al. (2016), where $a = b = 1/2$.

Lin et al. (2016) estimate truncated GGMs based on Hyvärinen's 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, such that no minimizer may exist. 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 that the loss is bounded and strongly convex and admits a unique minimizer. We apply this to the estimator for GGMs considered in Lin et al. (2016), which uses score-matching on $\mathbb{R}^m$, and to the *generalized score matching* estimator for pairwise interaction power models on $\mathbb{R}_+^m$ proposed in this paper. In these cases, we show, both empirically and theoretically, that the consistency results still hold (or even improve) if the positive values added are smaller than a threshold that is readily computable.

The rest of 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 normal distributions as an example. *Regularized generalized score matching* for graphical models is formulated in Section 4. The estimators for pairwise interaction power models are shown in Section 5, while theoretical consistency results are presented in Section 6, where we treat the probabilistically most tractable case of truncated GGMs. Simulation results and applications to RNAseq data are given in Section 7. Proofs for theorems in Sections 2–6 are presented in Appendices A and B. Additional experimental results are presented in Appendix C.

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 $\mathrm{vec}(\mathbf{K}) = (\kappa_{11}, \ldots, \kappa_{q1}, \kappa_{12}, \ldots, \kappa_{q2}, \ldots, \kappa_{1r}, \ldots, \kappa_{qr})^\top$. For a matrix $\mathbf{K} \in \mathbb{R}^{q\times q}$, $\mathrm{diag}(\mathbf{K}) \in \mathbb{R}^q$ denotes its diagonal, and for a vector $\boldsymbol{v} \in \mathbb{R}^q$, $\mathrm{diag}(\boldsymbol{v})$ 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

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

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

In this section, we review the original score matching and develop our generalized score matching estimators.

2.1. Original Score Matching

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

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x}. \tag{3}$$

The gradients in (3) 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, (3) can be rewritten as

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[\partial_{jj} \log p(\boldsymbol{x}) + \frac{(\partial_j \log p(\boldsymbol{x}))^2}{2} \right] \mathrm{d}\boldsymbol{x}, \tag{4}$$

plus a constant independent of p. The integral in (4) can be approximated by a sample average; this alleviates the need for 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 (3) and (4) 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*

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

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 distributions of interest, and assume every $P \in \mathcal{P}_+$ has a twice continuously differentiable density 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 continuously differentiable density. Let $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ be a.s. positive functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$, 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*

$$J_{\boldsymbol{h}}(P) = \int_{\mathbb{R}^m_+} \frac{1}{2} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} - \nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2}\|_2^2 \, \mathrm{d}\boldsymbol{x}, \tag{6}$$

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. 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, $p_1(\boldsymbol{x}) = \text{const} \times p_0(\boldsymbol{x})$ almost surely in $\mathbb{R}^m_+$. Since p_1 and p_0 are continuous densities supported on $\mathbb{R}^m_+$, 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 (5). 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 will consider the following assumptions:

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

(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 $\partial_j \log p(\boldsymbol{x}) \equiv \frac{\partial \log p(\boldsymbol{y})}{\partial y_j}\Big|_{\boldsymbol{y}=\boldsymbol{x}}$, $f(\boldsymbol{x}) \Big|_{x_j \searrow 0^+}^{x_j \nearrow +\infty} \equiv \lim_{x_j \nearrow +\infty} f(\boldsymbol{x}) - \lim_{x_j \searrow 0} f(\boldsymbol{x})$, "$\forall p \in \mathcal{P}_+$" 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, 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 and verified for exponential families.

Integration by parts yields the following theorem which shows that $J_{\boldsymbol{h}}$ from (6) is an expectation (under P_0) of a function that does not depend on p_0, similar to (4). The proof is given in Appendix A.1.

Theorem 3 *Under (A1) and (A2), the loss from (6) equals*

$$J_{\boldsymbol{h}}(P) = \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \Big[h_j'(x_j)\partial_j(\log p(\boldsymbol{x})) + h_j(x_j)\partial_{jj}(\log p(\boldsymbol{x})) + \frac{1}{2} h_j(x_j) \left(\partial_j(\log p(\boldsymbol{x}))\right)^2 \Big] \mathrm{d}\boldsymbol{x} \quad (7)$$

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 (7) as

$$\hat{J}_{\boldsymbol{h}}(P) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \Big\{ h_j'(X_j^{(i)})\partial_j(\log p(\boldsymbol{X}^{(i)})) + h_j(X_j^{(i)}) \left[\partial_{jj}(\log p(\boldsymbol{X}^{(i)})) + \frac{1}{2} \left(\partial_j(\log p(\boldsymbol{X}^{(i)})) \right)^2 \right] \Big\}. \quad (8)$$

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 $\hat{J}_{\boldsymbol{h}}(P)$. We note that this type of loss is also treated in slightly different settings in Parry (2016) and Almeida and Gidas (1993).

Remark 4 In the one-dimensional case, using the notation in Parry et al. (2012), $J_{\boldsymbol{h}}(P)$ and $\hat{J}_{\boldsymbol{h}}(P)$ correspond to $d(P_0, P)$ and $S(x, P)$, respectively, and can be generated by $\phi(x, p, p_1) \equiv -h(x)p_1^2/(2p)$ (c.f. Equations (39), (51), (53) and Section 10.1 therein). Thus Theorem 3 follows from this correspondence. While (A1) is equivalent to the condition implied by the boundary divergence $d_b = 0$ in that paper, (A2), which we assume for invoking Fubini-Tonelli due to multi-dimensionality, is not present. On the other hand, while Parry (2016) treats the multivariate case, it does not cover the connection between our $J_{\boldsymbol{h}}$ and $\hat{J}_{\boldsymbol{h}}$. Since ϕ is concave but not strictly concave in (p, p_1), the results in Parry (2016) only imply that P_0 is *a* minimizer, a weaker conclusion than Proposition 2.

3. Exponential Families

In this section, we study the case where $\mathcal{P}_+ \equiv \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ 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

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}_+^m, \tag{9}$$

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.s. 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 (8) can be rewritten as a quadratic function in $\boldsymbol{\theta} \in \mathbb{R}^r$:*

$$\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \text{const}, \quad \textit{where} \tag{10}$$

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}'_j(\boldsymbol{X}^{(i)})\boldsymbol{t}'_j(\boldsymbol{X}^{(i)})^\top \quad \textit{and} \tag{11}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[h_j(X_j^{(i)})b'_j(\boldsymbol{X}^{(i)})\boldsymbol{t}'_j(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}''_j(\boldsymbol{X}^{(i)}) + h'_j(X_j^{(i)})\boldsymbol{t}'_j(\boldsymbol{X}_i)\right] \tag{12}$$

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 a.s. invertible, and*

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

Then the minimizer of (10) is a.s. unique with closed-form solution $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$. Moreover,

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

Theorems 5 and 6 are proved in Appendix A.2. Theorem 5 clarifies the quadratic nature of the loss, and Theorem 6 provides a basis for asymptotically valid tests and confidence intervals for the parameter $\boldsymbol{\theta}$. Note that Condition (C1) holds if and only if $h_j(X_j) > 0$ a.s. and $[\boldsymbol{t}_j'(\boldsymbol{X}^{(1)}), \ldots, \boldsymbol{t}_j'(\boldsymbol{X}^{(n)})] \in \mathbb{R}^{r \times n}$ has rank r a.s. for some $j = 1, \ldots, m$.

The conclusion in Theorem 6 indicates that, similar to the estimator in Hyvärinen (2007) with $h_j(x) = x^2$, the closed-form solution for our generalized $\hat{\boldsymbol{\theta}}$ allows one to consistently estimate the canonical parameter in an exponential family distribution without needing to calculate the often complicated normalizing constant $\psi(\boldsymbol{\theta})$ or resort to numerical methods. Computational details are explicated in Section 5.3.

Below we illustrate the estimator $\hat{\boldsymbol{\theta}}$ in the case of univariate truncated normal distributions. We assume (A1)-(A2) and (C1)-(C2) throughout.

Example 3.1 *Univariate ($m = r = 1$) truncated normal distributions for mean parameter μ and variance parameter σ^2 have density*

$$p_{\mu,\sigma^2}(x) \propto \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}_+. \tag{13}$$

If σ^2 is known but μ unknown, then writing the density in canonical form as in (9) yields

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

Given an i.i.d. sample $X_1, \ldots, X_n \sim p_{\mu_0,\sigma^2}$, 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 $\lim_{x \searrow 0^+} h(x) = 0$, $\lim_{x \nearrow +\infty} h^2(x)(x-\mu_0)p_{\mu_0,\sigma^2}(x) = 0$ and the expectations are finite (for example, when $h(x) = o(\exp(Mx^2))$ for $M < \frac{1}{4\sigma^2}$), then

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

We recall that the Cramér-Rao lower bound *(i.e. the lower bound on the variance of any unbiased estimator) for estimating μ is*

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

Example 3.2 *Consider the univariate truncated normal distributions from (13) in the setting where the mean parameter μ is known but the variance parameter $\sigma^2 > 0$ is unknown. In canonical form as in (9), we write*

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

Given an i.i.d. sample $X_1, \ldots, X_n \sim p_{\mu,\sigma_0^2}$, 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 Example 3.1, $\lim_{x\nearrow+\infty} h^2(x)(x-\mu)^3 p_{\mu,\sigma_0^2}(x) = 0$, *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 7 In Example 3.2, 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.

In these examples, there is a benefit in using a bounded function h, which 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 (6), 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.

Figure 1 plots the asymptotic variance of $\hat{\mu}_h$ from Example 3.1, 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 2 is the analog of Figure 1 for $\hat{\sigma}_h^2$ from Example 3.2 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.

In both plots we order/color the curves based on their overall efficiency, so they have different colors in one from the other, although the same functions are presented. For all functions presented here (A1)–(A2) and (C1)–(C2) are satisfied.

4. Regularized Generalized Score Matching

In high-dimensional settings, when 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.

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 $\mathcal{P}_+ \equiv \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^r$. Adding an ℓ_1 penalty to (10), we obtain the regularized generalized score matching loss

$$\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1 \tag{14}$$

as in Lin et al. (2016). The loss in (14) involves a quadratic smooth part as in the familiar lasso loss for linear regression. However, although the matrix $\boldsymbol{\Gamma}$ is positive semidefinite,

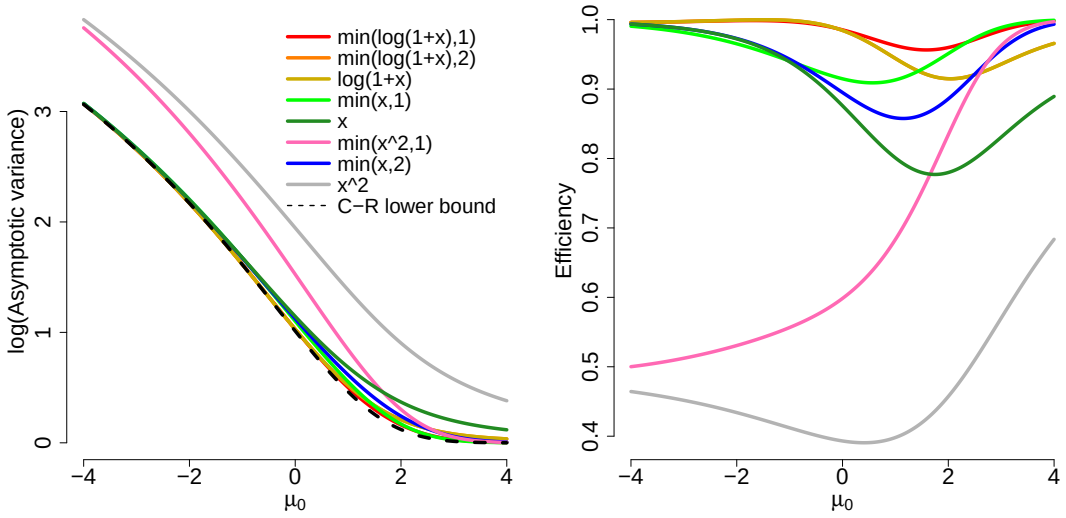


Figure 1: Log of asymptotic variance and efficiency with respect to the Cramér-Rao bound for $\hat{\mu}_h$ ($\sigma^2 = 1$ known).

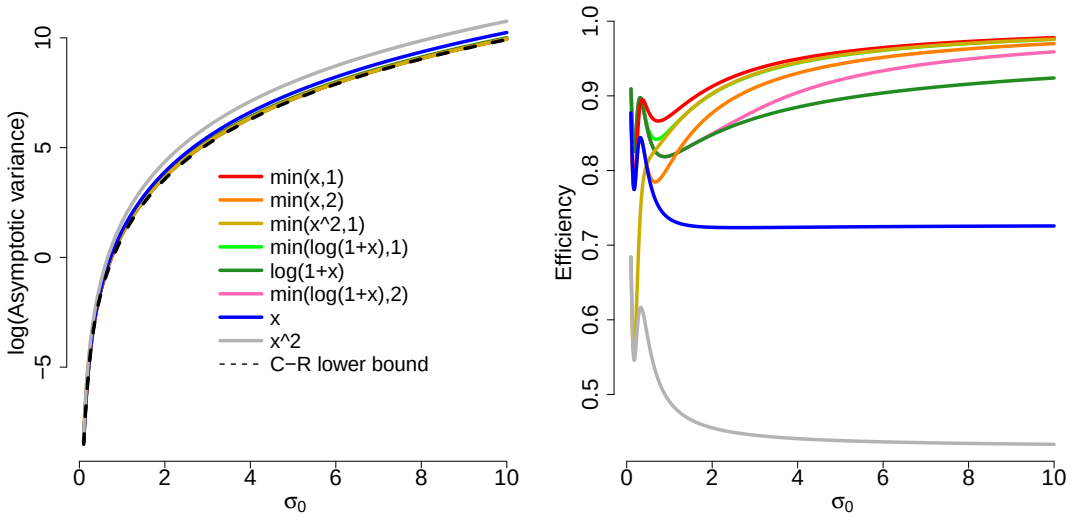


Figure 2: Log of asymptotic variance and efficiency with respect to the Cramér-Rao bound for $\hat{\sigma}_h^2$ ($\mu = 0.5$ known).

the regularized loss in (14) 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 (11), $\boldsymbol{\Gamma} = \mathbf{H}^\top \mathbf{H}$ 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

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

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 (14) 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 estimating the interaction matrix in pairwise models, $r = m^2$; thus, the condition $nm < r$ reduces to $n < m$, or $n < m + 1$ when both $\mathbf{K}$ and $\boldsymbol{\eta}$ are estimated.

To circumvent the unboundedness 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 $\lambda \geq 0$. For the special case of truncated GGMs, we will show that a result on consistent estimation holds if we choose $\gamma_\ell = \delta_0 \boldsymbol{\Gamma}_{\ell,\ell}$ for a suitably small constant $\delta_0 > 0$, 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 8 *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* amplifier $\boldsymbol{\gamma}$ *is the estimator*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma}}(\boldsymbol{\theta}) \equiv \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1. \tag{15}$$

In the case where $\boldsymbol{\gamma} = (\delta - 1)\mathrm{diag}(\boldsymbol{\Gamma})$ for some $\delta > 1$, we also call δ the *multiplier*. We note that $\hat{\boldsymbol{\theta}}$ from (15) is a *piecewise linear* function of λ (Lin et al., 2016).

5. Score Matching for Graphical Models for Non-negative Data

In this section we apply our generalized score matching estimator to a general class of graphical models for non-negative data.

5.1. A General Framework of Pairwise Interaction Models

We consider the class of pairwise interaction power models with density introduced in (1). We recall the form of the density:

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a + \boldsymbol{\eta}^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b}\right) \mathbb{1}_{\mathbb{R}_+^m}(\boldsymbol{x}), \tag{16}$$

where a and b are known constants, and the interaction matrix $\mathbf{K}$ and the vector $\boldsymbol{\eta}$ are parameters. When $b = 0$, we use the convention that $\frac{x^0-1}{0} \equiv \log x$ and apply the logarithm element-wise. Our focus will be on the interaction matrix $\mathbf{K}$ that determines the conditional independence graph through its support $S(\mathbf{K}) \equiv \{(i,j) : \kappa_{ij} \neq 0\}$. However, unless $\boldsymbol{\eta}$ is known or assumed to be zero, we also need to estimate $\boldsymbol{\eta}$ as a nuisance parameter. In the case where we assume $\boldsymbol{\eta} \equiv \mathbf{0}$ is known (i.e. the linear part $(\boldsymbol{x}^b - \mathbf{1}_m)/b$ is not present), we call the distribution (and the corresponding estimator) a *centered* distribution (estimator), in contrast to the general case termed *non-centered* when we assume $\boldsymbol{\eta} \neq \mathbf{0}$ or unknown.

We first give a set of sufficient conditions for the density to be valid, i.e., the right-hand side of (16) to be integrable. The proof is given in Appendix A.3.

Theorem 9 *Define conditions*

(CC1) $\mathbf{K}$ *is* strictly co-positive, *i.e.,* $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} > 0$ *for all* $\boldsymbol{v} \in \mathbb{R}^m_+ \backslash \{\mathbf{0}\}$*;*

(CC2) $2a > b > 0$*;*

(CC3) $a > 0$, $b = 0$, *and* $\eta_j > -1$ *for* $j = 1, \ldots, m$ *(*$\boldsymbol{\eta} \succ -\mathbf{1}_m$*).*

In the non-centered case, if (CC1) and one of (CC2) and (CC3) holds, then the function on the right-hand side of (16) is integrable over $\mathbb{R}^m_+$*. In the centered case, (CC1) and* $a > 0$ *are sufficient.*

We emphasize that (CC1) is a weaker condition than positive definiteness. Criteria for strict co-positivity are discussed in Väliaho (1986).

5.2. Implementation for Different Models

In this section we give some implementation details for the regularized generalized $\boldsymbol{h}$-score matching estimator defined in (15) applied to the pairwise interaction models from (16). We again let $\boldsymbol{\Psi} \equiv \left(\mathbf{K}^\top, \boldsymbol{\eta}\right)^\top \in \mathbb{R}^{(m+1)\times m}$. The unregularized loss is then

$$\hat{J}_{\boldsymbol{h}}(P) = \frac{1}{2}\mathrm{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}(\boldsymbol{x})\mathrm{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{\Psi}).$$

The general form of the matrix $\boldsymbol{\Gamma}$ and the vector $\boldsymbol{g}$ in the loss were given in equations (10)–(12). Here $\boldsymbol{\Gamma} \in \mathbb{R}^{(m+1)m\times(m+1)m}$ is block-diagonal, with the j-th $\mathbb{R}^{(m+1)\times(m+1)}$ block

$$\begin{aligned}
\boldsymbol{\Gamma}_j(\mathbf{x}) &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{11,j} & \boldsymbol{\Gamma}_{12,j} \\ \boldsymbol{\Gamma}_{12,j}^\top & \Gamma_{22,j} \end{bmatrix} \qquad (17)\\
&\equiv \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} h_j\left(X_j^{(i)}\right) X_j^{(i)2a-2} \boldsymbol{X}^{(i)a} \boldsymbol{X}^{(i)a\top} & -h_j\left(X_j^{(i)}\right) X_j^{(i)a+b-2} \boldsymbol{X}^{(i)a} \\ -h_j\left(X_j^{(i)}\right) X_j^{(i)a+b-2} \boldsymbol{X}^{(i)a\top} & h_j\left(X_j^{(i)}\right) X_j^{(i)2b-2} \end{bmatrix} \\
&= \frac{1}{n}\mathbf{y}^\top \mathbf{y}, \quad \mathbf{y} \equiv \left[-(\sqrt{\boldsymbol{h}_j(\boldsymbol{X}_j)} \circ \boldsymbol{X}_j^{a-1}) \circ \mathbf{x}^a \quad \sqrt{\boldsymbol{h}_j(\boldsymbol{X}_j)} \circ \boldsymbol{X}_j^{b-1}\right] \in \mathbb{R}^{n,m+1}, \qquad (18)
\end{aligned}$$

where the $\circ$ product between a vector and a matrix means an elementwise multiplication of the vector with each *column* of the matrix, and $\boldsymbol{h}_j(\boldsymbol{X}_j) \equiv [h_j(X_j^{(1)}), \ldots, h_j(X_j^{(n)})^\top] \in \mathbb{R}^m$.

Furthermore, $\boldsymbol{g} \equiv \begin{bmatrix} \mathrm{vec}(\mathbf{g}_1) \\ \boldsymbol{g}_2 \end{bmatrix} \in \mathbb{R}^{(m+1)m}$, where $\mathbf{g}_1$ and $\boldsymbol{g}_2$ correspond to each entry of $\mathbf{K}$ and $\boldsymbol{\eta}$, respectively. The j-th column of $\mathbf{g}_1 \in \mathbb{R}^{m \times m}$, written as $\boldsymbol{g}_{1,j}(\mathbf{x})$, is

$$\frac{1}{n}\sum_{i=1}^{n}\left(h_j'\left(X_j^{(i)}\right){X_j^{(i)}}^{a-1} + (a-1)h_j\left(X_j^{(i)}\right){X_j^{(i)}}^{a-2}\right)\boldsymbol{X}^{(i)a} + ah_j\left(X_j^{(i)}\right){X_j^{(i)}}^{2a-2}\boldsymbol{e}_{j,m},$$

where $\boldsymbol{e}_{j,m}$ is the m-vector with 1 at the j-th position and 0 elsewhere, and the j-th entry of $\boldsymbol{g}_2 \in \mathbb{R}^m$ is

$$g_{2,j} = \frac{1}{n}\sum_{i=1}^{n} -h_j'\left(X_j^{(i)}\right){X_j^{(i)}}^{b-1} - (b-1)h_j\left(X_j^{(i)}\right){X_j^{(i)}}^{b-2}.$$

These formulae also hold for $b = 0$ since $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ only depend on the gradient of the log density, and $\frac{\mathrm{d}(x^b-1)/b}{\mathrm{d}x} = x^{b-1}$ also holds for $b = 0$. In the centered case where we know $\boldsymbol{\eta}_0 \equiv \mathbf{0}$, we only estimate $\mathbf{K} \in \mathbb{R}^{m \times m}$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{m^2 \times m^2}$ is still block-diagonal, with the j-th block being the $\boldsymbol{\Gamma}_{11,j}$ submatrix in (17), while $\boldsymbol{g}$ is just $\mathrm{vec}(\mathbf{g}_1)$. Since b only appears in the $\boldsymbol{\eta}$ part of the density, the formulae only depend on a in the centered case.

We emphasize that it is indeed necessary to introduce amplifiers $\boldsymbol{\gamma} \succ \mathbf{0}$ or a multiplier $\delta > 1$ in addition to the ℓ_1 penalty. It is clear from (18) that $\mathrm{rank}(\boldsymbol{\Gamma}_j) \leq \min\{n, m+1\}$ (or $\min\{n, m\}$ if centered). Thus, $\boldsymbol{\Gamma}$ is non-invertible when $n \leq m$ (or $n < m$ if centered) and $\boldsymbol{g}$ need not lie in its column span.

We claim that including amplifiers/multipliers for the submatrices $\boldsymbol{\Gamma}_{11,j}$ only 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 $\boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}}$ be our amplified version of the matrix $\boldsymbol{\Gamma}_j$ from (21), so

$$\boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}} \;=\; \begin{pmatrix} \boldsymbol{\Gamma}_{11,j} + \mathrm{diag}(\gamma_1, \ldots, \gamma_m) & \boldsymbol{\Gamma}_{12,j} \\ \boldsymbol{\Gamma}_{12,j}^\top & \Gamma_{22,j} \end{pmatrix}.$$

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

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

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

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

almost surely; recall that $\Gamma_{22,j} = \frac{1}{n}\sum_{i=1}^{n} h_j\left(X_j^{(i)}\right) X_j^{2b-2}$. We conclude that $\boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}}$ (and thus the entire amplified $\boldsymbol{\Gamma}$) is a.s. positive definite, which ensures unique existence of the loss minimizer.

Given the formulae for $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$, one adds the ℓ_1 penalty on $\boldsymbol{\Psi}$ to get the regularized loss (24). Our methodology readily accommodates two different choices of the penalty parameter λ for $\mathbf{K}$ and $\boldsymbol{\eta}$. This is also theoretically supported for truncated GGMs, since if the ratio of the respective values $\lambda_{\mathbf{K}}$ and $\lambda_{\boldsymbol{\eta}}$ is fixed, the proof of the theorems in Section 6 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 $\hat{\boldsymbol{\eta}} \equiv \boldsymbol{\Gamma}_{22}^{-1}\left(\boldsymbol{g}_2 - \boldsymbol{\Gamma}_{12}^{\top}\mathrm{vec}(\hat{\mathbf{K}})\right)$, with $\hat{\mathbf{K}}$ the minimizer of the profiled loss

$$\hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma},\mathrm{profile}}(\mathbf{K}) \equiv \frac{1}{2}\mathrm{vec}(\mathbf{K})^{\top}\boldsymbol{\Gamma}_{\boldsymbol{\gamma},11.2}\mathrm{vec}(\mathbf{K}) - (\boldsymbol{g}_1 - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{g}_2)^{\top}\mathrm{vec}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \tag{19}$$

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, however. 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.

Example 5.3 *The truncated normal model comprises the density*

$$p_{\boldsymbol{\mu},\mathbf{K}}(\boldsymbol{x}) \;\propto\; \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\mathbf{K}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\mathbb{1}_{[0,\infty)^m}(\boldsymbol{x}). \tag{20}$$

This corresponds to (16) with $a = b = 1$, and $\boldsymbol{\eta} = \mathbf{K}\boldsymbol{\mu}$. The j-th $(m+1)\times(m+1)$ block of $\boldsymbol{\Gamma}(\mathbf{x})$ is

$$\frac{1}{n}\begin{bmatrix} \mathbf{x}^{\top}\mathrm{diag}(\boldsymbol{h}_j(\boldsymbol{X}_j))\mathbf{x} & -\mathbf{x}^{\top}\boldsymbol{h}_j(\boldsymbol{X}_j) \\ -\boldsymbol{h}_j(\boldsymbol{X}_j)^{\top}\mathbf{x} & \boldsymbol{h}_j(\boldsymbol{X}_j)^{\top}\mathbf{1}_n \end{bmatrix}. \tag{21}$$

Partitioning 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 $\boldsymbol{\Psi}_j$, the k-th entry of $\boldsymbol{g}_j(\mathbf{x})$ is

$$g_{jk}(\mathbf{x}) \equiv \begin{cases} \frac{1}{n}\sum_{i=1}^n h_j'\left(X_j^{(i)}\right)X_k^{(i)} & \text{if } k \leq m,\ k \neq j, \\ \frac{1}{n}\sum_{i=1}^n h_j'\left(X_j^{(i)}\right)X_k^{(i)} + h_j\left(X_j^{(i)}\right) & \text{if } k = j, \\ -\frac{1}{n}\sum_{i=1}^n h_j'\left(X_j^{(i)}\right) & \text{if } k = m+1. \end{cases} \tag{22}$$

Example 5.4 *The exponential square-root graphical model in Inouye et al. (2016) has*

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\sqrt{\boldsymbol{x}}^{\top}\mathbf{K}\sqrt{\boldsymbol{x}} + 2\boldsymbol{\eta}^{\top}\sqrt{\boldsymbol{x}}\right)\mathbb{1}_{[0,\infty)^m}(\boldsymbol{x}),$$

which corresponds to (16) with $a = b = 1/2$. We refer to this as the exponential *model. In this case, the j-th $\mathbb{R}^{(m+1)\times(m+1)}$ block of $\boldsymbol{\Gamma}$ is*

$$\boldsymbol{\Gamma}_j(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^n \frac{h_j\left(X_j^{(i)}\right)}{X_j^{(i)}}\begin{pmatrix} -\sqrt{\boldsymbol{X}^{(i)}} \\ 1 \end{pmatrix}\left(-\sqrt{\boldsymbol{X}^{(i)}}^{\top}, 1\right)$$

and $\boldsymbol{g} = \mathrm{vec}(\mathbf{g}_0)$, where the j-th column of $\mathbf{g}_0 \in \mathbb{R}^{(m+1)\times m}$ is

$$\boldsymbol{g}_j(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^n \frac{2h_j'\left(X_j^{(i)}\right)X_j^{(i)} - h_j\left(X_j^{(i)}\right)}{2{X_j^{(i)}}^{3/2}}\begin{pmatrix} \sqrt{\boldsymbol{X}^{(i)}} \\ -1 \end{pmatrix} + \frac{h_j\left(X_j^{(i)}\right)}{2X_j^{(i)}}\boldsymbol{e}_{j,m+1}.$$

Example 5.5 *If $a = 1/2$ and $b = 0$, then (16) becomes*

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\sqrt{\boldsymbol{x}}^\top \mathbf{K}\sqrt{\boldsymbol{x}} + \boldsymbol{\eta}^\top \log(\boldsymbol{x})\right) \mathbb{1}_{(0,\infty)^m}(\boldsymbol{x}). \tag{23}$$

If $\mathbf{K}$ is diagonal in this case, then $\boldsymbol{X} \sim p_{\boldsymbol{\eta},\mathbf{K}}$ has independent entries with X_j following the gamma distribution with rate κ_{jj} and shape $\eta_j + 1$, which gives an intuition for condition (CC3) $\eta_j > -1$ in Theorem 9. We can thus view (23) as a multivariate gamma distribution with pairwise interactions among the covariates, and call this the gamma *model. For this model, the j-th block of $\boldsymbol{\Gamma}$ is*

$$\boldsymbol{\Gamma}_j(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{h_j\left(X_j^{(i)}\right)}{{X_j^{(i)}}^2} \begin{pmatrix} -\sqrt{X_j^{(i)}\boldsymbol{X}^{(i)}} \\ 1 \end{pmatrix} \left(-\sqrt{X_j^{(i)}\boldsymbol{X}^{(i)}}^\top, 1\right)$$

and the part of $\boldsymbol{g}$ corresponding to $\mathbf{K}_j$ is

$$\boldsymbol{g}_{1,j}(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{2h_j'\left(X_j^{(i)}\right)X_j^{(i)} - h_j\left(X_j^{(i)}\right)}{2{X_j^{(i)}}^{3/2}}\sqrt{\boldsymbol{X}^{(i)}} + \frac{h_j\left(X_j^{(i)}\right)}{2X_j^{(i)}}\boldsymbol{e}_{j,m},$$

while the part for η_j is

$$g_{2.j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{h_j\left(X_j^{(i)}\right)}{{X_j^{(i)}}^2} - \frac{h_j'\left(X_j^{(i)}\right)}{X_j^{(i)}}.$$

We note that the $\boldsymbol{\Gamma}_{11,j}$ sub-matrix of $\boldsymbol{\Gamma}_j$ and the $\boldsymbol{g}_{1,j}$ sub-vector of $\boldsymbol{g}_j$ for the gamma model are the same as those for the exponential model, since $a = 1/2$ in both cases and the parts involving $\mathbf{K}$ in the densities are the same.

5.3. Computational Details

In the most general exponential family setting, as in Eq. (10)–(12) in Theorem 5, the time complexity for forming $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{g} \in \mathbb{R}^r$ is $\mathcal{O}\left(nm(f_{b'}(m) + r^2 + r(f_{t'}(m) + f_{t''}(m)))\right)$. Here $f_{b'}(m)$ is the average time complexity for calculating $\partial_j b(\boldsymbol{x})$ over $j = 1,\ldots,m$, and similarly $f_{t'}(m)$ for $\partial_j t_\ell(\boldsymbol{x})$ and $f_{t''}(m)$ for $\partial_{jj} t_\ell(\boldsymbol{x})$ over $j = 1,\ldots,m$ and $\ell = 1\ldots,r$. In many applications, however, these three functions would be constant in m, thus giving an $\mathcal{O}(nmr^2)$ computational complexity, with the dominating term coming from the operations for $\boldsymbol{t}_j'\boldsymbol{t}_j'^\top$ in $\boldsymbol{\Gamma}$ since $\boldsymbol{\Gamma}$ is of dimension $r \times r$.

For pairwise interaction power models, $r = m^2$ and the formula above becomes $\mathcal{O}(nm^5)$. However, since $\boldsymbol{\Gamma}$ is block-diagonal with only m^3 nonzero entries and by the special form of $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}^a\boldsymbol{x}^{a\top}$, the true complexity is in fact $\mathcal{O}(nm^3)$.

While the introduction of the ℓ_1 penalty inevitably precludes the estimator from having a closed-form solution and introduces non-differentiability, state-of-art numerical optimization algorithms, such as coordinate-descent (Friedman et al., 2007), can be applied for fast estimation. To speed up estimation, one can usually use warm starts using the solution from the previous λ's, as well as lasso-type strong screening rules (Tibshirani et al., 2012) to eliminate components of $\hat{\boldsymbol{\theta}}$ that are known a priori to have zero estimates.

In our implementation for pairwise interaction models of Section 5.1 (that will become available in an R package), we optimize our loss functions with respect to a symmetric matrix $\hat{\mathbf{K}}$; in the non-centered case the vector $\hat{\boldsymbol{\eta}}$ is also included. We use a coordinate-descent method analogous to Algorithm 2 in Lin et al. (2016), where in each step we update each element of $\hat{\mathbf{K}}$ and $\hat{\boldsymbol{\eta}}$ based on the other entries from the previous steps, while maintaining symmetry. In our simulations in Section 7 we always scale the data matrix by column ℓ_2 norms before proceeding to estimation. Note that estimation of $\hat{\mathbf{K}}$ without symmetry can be parallelized as the loss can be decomposed into a sum over the columns.

5.4. Choice of the Function $\boldsymbol{h}$

In this subsection we discuss the requirements on the function $\boldsymbol{h}$ as well as some reasonable choices of $\boldsymbol{h}$.

5.4.1. Requirements on $\boldsymbol{h}$

In Section 2.2, we presented two assumptions (A1) and (A2) under which the generalized score-matching loss is valid, i.e., the integration by parts is justified and Theorem 3 holds. In this section, we present some sufficient (and nearly necessary) requirements on $\boldsymbol{h}$ such that (A1) and (A2) are satisfied.

Definition 10 *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 $\boldsymbol{h} \in \mathcal{H}_{a,b}$ (for simplicity we omit the dependency on m) if for all $j = 1, \ldots, m$:*

i) h_j is absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$, and thus has derivative h_j' a.s.;

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

iii) h_j and h_j' are both bounded by some piecewise powers of x a.s. on $\mathbb{R}_+$;

iv) $\lim_{x \searrow 0+} h_j(x)/x_j^q = 0$, where $q \equiv \begin{cases} \max\{1-a, 1-b\} & \text{if } b > 0, \\ 1 - \eta_{0,j} & \text{if } b = 0. \end{cases}$

Theorem 11 *Assume every P in the family of distribution $\mathcal{P}_+$ satisfies (CC1)–(CC3) and thus has finite normalizing constants. If $\boldsymbol{h} \in \mathcal{H}_{a,b}$, then (A1) and (A2) are satisfied.*

In centered models, where $\boldsymbol{\eta} \equiv \mathbf{0}$, we can assume $b = 2a$ and iv) in the definition of $\mathcal{H}_{a,2a}$ has $q = 1 - a$. For truncated GGMs, $a = b = 1$, so iv) in Definition 10 is simply $\lim_{x_j \searrow 0^+} h_j(x_j) = 0$.

In the case of $b = 0$, $\boldsymbol{\eta}$ is an unknown parameter, and (CC3) requires each of its component to be greater than -1. If one has prior information on $\boldsymbol{\eta}$ or restricts the parameter space for $\boldsymbol{\eta}$, the requirement reduces to $h_j(x_j) = o(x_j^{1-\eta_{0,j}})$ as $x_j \searrow 0^+$. Otherwise, it suffices to require $h_j(x_j) = o(x_j^2)$. Note that this is only a condition for $x_j \searrow 0^+$, and the globally quadratic behavior of $h_j(x_j) = x_j^2$ from the original score matching is not needed on the entire $\mathbb{R}_+$, leaving opportunities for improvements.

5.4.2. Reasonable Choices of $\boldsymbol{h}$

Assume a common univariate h for all components in $\boldsymbol{h}$. Inspired by Theorem 11, we consider h that behaves like a power of x both as $x \nearrow +\infty$ and as $x \searrow 0^+$. Since the requirements on the two tails are separate, we can choose h to be a piecewise defined function that joins two powers with possibly different degrees. In other words, $h(x) = \min(x^{p_1}, cx^{p_2})$ for some powers $p_1 \geq p_2 \geq 0$ and constant $c > 0$. Only one constant c is required since generalized score matching is invariant to scaling of h. In determining the exact power of p_1 we have the following considerations:

a) In the centered case:

   (i) (A1) and (A2): Theorem 11 requires that $p_1 \geq 1 - a$.

   (ii) "Controlled $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ for $\boldsymbol{x}^a$": We propose avoiding poles at the origin for the entries of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$. The formula for $\boldsymbol{\Gamma}_{11}$ in (18) shows that to this end $\sqrt{h(x)}x^{a-1}$ needs to have a non-negative degree . This requires $p_1 \geq 2 - 2a$. The formula for $\boldsymbol{g}_1$ similarly shows that $h'(x)x^{a-1}$, $h(x)x^{a-2}$ and $h(x)x^{2a-2}$ all need to have a non-negative degree for small x. This requires $p_1 \geq 2 - a$.

b) In the non-centered case, in addition to (i) and (ii),

   (iii) (A1) and (A2): Theorem 11 requires $p_1 \geq \max\{1 - a, 1 - b\}$ for $b > 0$, or $1 - \min_j \eta_{0,j}$ for $b = 0$.

   (iv) "Controlled $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ for $\boldsymbol{x}^b$": From the definition of $\boldsymbol{\Gamma}_{22}$ and $\boldsymbol{g}_2$ and by the same reasoning as above, $\sqrt{h(x)}x^{b-1}$, $h'(x)x^{b-1}$ and $h(x)x^{b-2}$ need to be non-negative powers of x, thus requiring $p_1 \geq \max\{2 - b, 2 - 2b\} = 2 - b$.

The choice of p_2, is only relevant for large data points. Our main consideration is then merely how well $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ concentrate on their true population values (Theorem 13). From this perspective, our intuition is that p_2 should be chosen small so that the tails of the distributions of the entries of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are well-behaved. Thus, we can choose $p_2 = 0$, in which case $h(x) = \min(x^{p_1}, c)$ is a truncated power.

5.5. Tuning Parameter Selection

By treating the unpenalized loss (i.e., $\lambda = 0$, $\boldsymbol{\gamma} = 0$) as a 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 $\hat{S}^\lambda \equiv \{(i,j) : \hat{\kappa}^\lambda_{ij} \neq 0, i < j\}$, 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 $\hat{S}^\lambda$.

We use the eBIC instead of the ordinary BIC (Bayesian Information Criterion) since the BIC tends to choose an overly complex model when the model space is large, as encountered in the high-dimensional setting. The extension in eBIC comes from the last

term in the above display which can be motivated by a prior distribution under which the number of edges in the conditional independence graph is uniformly distributed; see also Żak-Szatkowska and Bogdan (2011) and Barber and Drton (2015).

6. Theory for Graphical Models

In our regularized generalized score matching framework, we introduced the amplifiers/ multipliers to address the inexistence problem. We also proposed using a general function $\boldsymbol{h}$ in place of $\boldsymbol{x}^2$ as a means to improve estimation accuracy. This section provides a theoretical analysis of these two aspects.

In Section 6.1, we present the theory for our regularized generalized score matching estimators for general pairwise interaction models before going into the details for the special cases of (truncated) GGMs. Next, we show that a specific choice of amplifiers/multipliers yields consistent estimation without the need for tuning. This point is important even in the case of Gaussian models on all of $\mathbb{R}^m$. Therefore, in Section 6.2 we digress from non-negative data and consider the original score matching of Hyvärinen (2005) for centered Gaussian distributions. Finally, in Section 6.3, we derive probabilistic results for $\hat{\boldsymbol{\Psi}}$ based on Theorem 13, justifying the benefits of using a general bounded $\boldsymbol{h}$ over $\boldsymbol{x}^2$ in the non-negative setting. As the most important models from the class of pairwise interaction power models over $\mathbb{R}^m_+$, we only treat truncated GGMs since they have the most tractable concentration bounds; this case also provides a comparison to Corollary 2 in Lin et al. (2016), which uses $\boldsymbol{x}^2$.

6.1. Theory for Pairwise Interaction Models

The graphical models we treat are parametrized by the interaction matrix $\mathbf{K}$ and the coefficients $\boldsymbol{\eta}$ on $(\boldsymbol{x}^b - \mathbf{1}_m)/b$. 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 specify our regularized $\boldsymbol{h}$-score matching loss as

$$\hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma}}(\boldsymbol{\Psi}) \equiv \underset{\boldsymbol{\Psi}\in\mathbb{R}^{r_1\times r_2}}{\operatorname{argmin}} \frac{1}{2}\operatorname{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x})\operatorname{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}(\mathbf{x})^\top \operatorname{vec}(\boldsymbol{\Psi}) + \lambda\|\boldsymbol{\Psi}\|_1. \tag{24}$$

In the non-centered case we thus take $\boldsymbol{\Psi} = [\mathbf{K}, \boldsymbol{\eta}]^\top \in \mathbb{R}^{m(m+1)\times m}$. In the centered case, $\boldsymbol{\Psi}$ is simply the $m \times m$ interaction matrix $\mathbf{K}$. Following related prior work such as Lin et al. (2016), for ease of proof we allow the matrix $\mathbf{K}$ to be nonsymmetric, which allows us to decouple optimization over the different columns of $\mathbf{K}$ or $\boldsymbol{\Psi}$, while in our implementations we ensure that $\mathbf{K}$ is symmetric.

Definition 12 *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*

$$c_{\boldsymbol{\Gamma}_0} \equiv |||(\boldsymbol{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty}. \tag{25}$$

Finally, $\boldsymbol{\Gamma}_0$ satisfies the irrepresentability condition with incoherence parameter $\alpha \in (0,1]$ and edge set S_0 *if*

$$|||\boldsymbol{\Gamma}_{0,S_0^cS_0}(\boldsymbol{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty} \leq (1-\alpha). \tag{26}$$

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 13 *Suppose $\boldsymbol{\Gamma}_0$ has $\boldsymbol{\Gamma}_{0,S_0S_0}$ invertible and satisfies the irrepresentability condition (26) with incoherence parameter $\alpha \in (0,1]$. Assume that*

$$\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty < \epsilon_1, \qquad \|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2, \tag{27}$$

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 (24) 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 16 and 17 in Section 6.3. Before going into these examples, we state a general corollary.

Corollary 14 *Under the assumptions of Theorem 13, the matrix $\hat{\boldsymbol{\Psi}}$ minimizing (24) satisfies*

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

6.2. Revisiting Gaussian Score Matching

In this section we consider estimating the inverse covariance matrix $\mathbf{K}$ of a centered Gaussian distribution $\mathrm{N}(\mathbf{0}, \mathbf{K})$, which of course has density proportional to (2) on all of $\mathbb{R}^m$. As shown, e.g., in Example 1 of Lin et al. (2016), the ℓ_1-regularized score matching loss then takes the form

$$\frac{1}{2}\operatorname{tr}(\mathbf{K}\mathbf{K}\mathbf{x}\mathbf{x}^\top) - \operatorname{tr}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \tag{28}$$

which can be written as (14) with $\boldsymbol{\theta} = \operatorname{vec}(\mathbf{K})$, $\boldsymbol{\Gamma} = \operatorname{diag}(\mathbf{x}\mathbf{x}^\top, \ldots, \mathbf{x}\mathbf{x}^\top)$ and $\boldsymbol{g} = \operatorname{vec}(\mathbf{I}_m)$. Thus, in general, 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. Hence, an amplifier/multiplier on the diagonals of $\boldsymbol{\Gamma}$ is needed. We have the following theorem on the estimator using the amplification.

Theorem 15 *Suppose the data matrix* $\mathbf{x}$ *holds* n *i.i.d. copies of* $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}_0)$. *Adopt the amplifying in Section 4 and redefine the loss in (28) as*

$$\frac{1}{2}\mathrm{tr}(\mathbf{KKG}) - \mathrm{tr}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \quad \mathbf{G}_{jk} = (\mathbf{x}\mathbf{x}^\top)_{jk}\left(\mathbb{1}_{\{j\neq k\}} + (\delta-1)\mathbb{1}_{\{j=k\}}\right), \tag{29}$$

where $1 < \delta < 2 - \left(1 + 80\sqrt{\log m/n}\right)^{-1}$. *Let* $\hat{\mathbf{K}}$ *be the resulting estimator. Let* $c^* \equiv 12800\,(\max_j \boldsymbol{\Sigma}_{0,jj})^2$ *and* $c_1 = 4c_{\boldsymbol{\Gamma}_0}/\alpha$. *If for some* $\tau > 2$, *the regularization parameter and the sample size satisfy*

$$\lambda > (2c_{\mathbf{K}_0}(2-\alpha)\sqrt{c^*(\tau\log m + \log 4)/n})/\alpha,$$
$$n > \max(c^* c_1^2 d_{\mathbf{K}_0}^2, 2)(\tau\log m + \log 4),$$

then $\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda$ *with probability* $1 - m^{2-\tau}$.

In Corollary 1 of Lin et al. (2016) the same results were shown with $c^* \equiv 3200\,(\max_j \boldsymbol{\Sigma}_{0,jj})^2$ when a unique minimizer exists, but the existence was not guaranteed.

6.3. Generalized Score Matching for Truncated GGMs

Next, we provide theory for the regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\boldsymbol{\Psi}}$ in the special case of truncated GGMs. Again, assume a common h for all components in $\boldsymbol{h}$.

Theorem 16 *Suppose the data matrix* $\mathbf{x}$ *holds* n *i.i.d. copies of* $\boldsymbol{X} \sim \mathrm{TN}(\mathbf{0}, \mathbf{K}_0)$, *where the mean parameter is known to be zero. Assume that* $\boldsymbol{h} \in \mathcal{H}_{1,1}$ *and that* $0 \leq h \leq M$, $0 \leq h' \leq M'$ *a.s. for constants* M, M', *and choose* $\boldsymbol{\gamma} = (\delta-1)\mathrm{diag}(\boldsymbol{\Gamma})$ *with*

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

Suppose that the $\boldsymbol{\Gamma}_{0,S_0S_0}$ *block of* $\boldsymbol{\Gamma}_0$ *is invertible and* $\boldsymbol{\Gamma}_0$ *satisfies the irrepresentability condition (26) with* $\alpha \in (0,1]$ *and true edge set* S_0. *Define* $c_{\boldsymbol{X}} \equiv 2\max_j \left(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\,\mathbb{E}_0 X_j\right)$. *If for* $\tau > 3$ *the sample size and the regularization parameter satisfy*

$$n > \mathcal{O}\left(\tau\log m \max\left\{\frac{M^2 c_{\boldsymbol{\Gamma}_0}^2 c_{\boldsymbol{X}}^4 d_{\mathbf{K}_0}^2}{\alpha^2}, \frac{M c_{\boldsymbol{\Gamma}_0} c_{\boldsymbol{X}}^2 d_{\mathbf{K}_0}}{\alpha}\right\}\right), \tag{30}$$

$$\lambda > \mathcal{O}\left[(M c_{\mathbf{K}_0} c_{\boldsymbol{X}}^2 + M' c_{\boldsymbol{X}} + M)\left(\sqrt{\frac{\tau\log m}{n}} + \frac{\tau\log m}{n}\right)\right], \tag{31}$$

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 (24) is unique, has its support included in the true support,* $\hat{S} \equiv S(\hat{\mathbf{K}}) \subseteq S_0$, *and satisfies*

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

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

where $c_{\mathbf{\Gamma}_0}$ is defined in (25).

(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 $\mathrm{sign}(\hat{\kappa}_{jk}) = \mathrm{sign}(\kappa_{0,jk})$ for all $(j,k) \in S_0$.

The theorem is proved in Appendix A.4, 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 calibrates 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 $n \ge \mathcal{O}(d_{\mathbf{K}_0}^2(\log m^\tau)^8)$ in Lin et al. (2016).

For the non-centered case, by definition, $c_{\mathbf{\Psi}_0} \equiv |||\mathbf{\Psi_0}^\top|||_{\infty,\infty} \le c_{\mathbf{K}_0} + \|\boldsymbol{\eta}_0\|_\infty$, $d_{\mathbf{\Psi}_0} \le d_{\mathbf{K}_0} + 1$. The proof given for Theorem 16 goes through again here, and we have the following consistency results.

Theorem 17 *Suppose the data matrix holds n i.i.d. copies of $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}_0, \mathbf{K}_0)$. Assume that $\boldsymbol{h} \in \mathcal{H}_{1,1}$ and that $0 \le h \le M$, $0 \le h' \le M'$ a.s. for constants M, M'. Let $\boldsymbol{\gamma}$ be a vector of amplifiers that are non-zero only for the diagonal entries of the matrices $\mathbf{\Gamma}_{11,j}$, amplifying those by $(\delta-1)\mathrm{diag}(\mathbf{\Gamma}_{11,j})$ with*

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

Suppose further that $\mathbf{\Gamma}_{0,S_0S_0}$ is invertible and satisfies the irrepresentability condition (26) with $\alpha \in (0,1]$. Define $c_{\boldsymbol{X}} \equiv 2\max_j\left(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\,\mathbb{E}_0 X_j\right)$. Suppose for $\tau > 3$ the sample size and the regularization parameter satisfy

$$n > \mathcal{O}\left(\tau\log m\max\left\{\frac{M^2 c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0}^2 c_{\boldsymbol{X}}^4 d_{\mathbf{\Psi}_0}^2}{\alpha^2}, \frac{M c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0} c_{\boldsymbol{X}}^2 d_{\mathbf{\Psi}_0}}{\alpha}\right\}\right), \tag{32}$$

$$\lambda > \mathcal{O}\left[(M c_{\mathbf{\Psi}_0} c_{\boldsymbol{X}}^2 + M' c_{\boldsymbol{X}} + M)\left(\sqrt{\frac{\tau\log m}{n}} + \frac{\tau\log m}{n}\right)\right], \tag{33}$$

where $c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0}$ is $c_{\mathbf{\Gamma}_0}$ as in (25) but with notation $\mathbf{\Psi}_0$ to differentiate it from the centered case. Then the following statements hold with probability $1 - m^{3-\tau}$:

(a) The regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\mathbf{\Psi}}$ that minimizes (24) is unique, has its support included in the true support, $\hat{S} \equiv S(\hat{\mathbf{\Psi}}) \subseteq S_0$, and satisfies

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

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

(b) Moreover, if

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

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

Remark 18 The quantity $c_{\boldsymbol{X}}$ in Theorem 17 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 $(\mathbf{K}_0)^{-1}_{jj}$, then $c_{\boldsymbol{X}}$ seems to scale as $\boldsymbol{\mu}_0$, which negatively impacts the guarantees stated in Theorem 17. However, as in the one-dimensional case for estimation of μ_0 (Example 3.1), our estimator should automatically adapt to the large mean parameter. This suggests that it might be possible to improve our analysis involving $c_{\boldsymbol{X}}$.

7. Numerical Experiments

In this section, we compare the performance of our estimator with different choices of $\boldsymbol{h}$ to the existing approaches for pairwise interaction power models. In our simulation experiments, we consider $m = 100$ variables and $n = 80$ and $n = 1000$ samples, corresponding to high- and low-dimensional settings. We also tried intermediate sample sizes between these two extremes, but found no interesting result worth reporting. For $n = 80$, amplification is necessary. Except in Section 7.2.2, the amplifier is set based on Theorem 16 to $\delta = C(n, m) = 1.8647$ for truncated GGMs. The same amplifier is also used for settings with other a and b. For $n = 1000$, we consider $\delta = 1$, i.e., no amplification, and $\delta = C(n, m) = 1.6438$ (again, based on Theorem 16). Throughout, we assume a common univariate h for all components in $\boldsymbol{h}$.

7.1. Structure of K

The underlying interaction matrices are selected as follows: 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. Thus, $\mathbf{K}_0$ is block-diagonal. In each block, each lower-triangular element is set to 0 with probability $1-\pi$ for some $\pi \in (0, 1)$, and is otherwise drawn from Uniform$[0.5, 1]$. 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.

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 $\pi = 0.8$, which is in accordance with Lin et al. (2016). For $n = 80$, we set $\pi = 0.2$. This way $n/(d^2_{\mathbf{K}_0} \log m)$ is roughly constant; recall Theorems 16 and 17 for truncated GGMs.

In Appendix C, we report results on *Erdös-Rényi graphs*, which lead to similar conclusions.

7.2. Truncated GGMs

Given our focus on truncated GGMs and their relevance in graphical modeling applications, we start with experiments for these models.

7.2.1. Choice of $\boldsymbol{h}$

Our estimator requires choosing a function $\boldsymbol{h} : \mathbb{R}_+^m \to \mathbb{R}_+^m$. For simplicity, we will always specify $\boldsymbol{h}(\boldsymbol{x}) = (h(x_1), \ldots, h(x_m))$ for a single non-decreasing univariate function $h : \mathbb{R}_+ \to \mathbb{R}_+$, i.e. all coordinates share the same h function.

As previously explained, $\boldsymbol{h} \in \mathcal{H}_{a,b}$ is a sufficient condition for assumptions (A1)-(A2), as well as (C1)-(C2) in the case of unregularized estimators. Only in the proofs of our theoretical guarantees in Section 6 for truncated GGMs, did we require h to be bounded and to have bounded derivatives. As motivated by the discussion in Section 5.4.2, we consider truncated and untruncated powers, $\min(x, c)$ and x (since $2 - a = 2 - b = 1$); we evaluate this choice by contrasting them with powers $x^{1.5}$ and x^2. We also explore functions like $\log(1 + x)$ that seem natural and are linear near 0. In particular, we make a further comparison to functions linear near 0 with a finite asymptote as $x \nearrow +\infty$ but differentiable everywhere: MCP- (Fan and Li, 2001) and SCAD-like (Zhang, 2010) functions defined below. The results we report are based on selections of best performing choices of h.

$$\mathrm{SCAD}(x; \lambda, \gamma) \equiv \begin{cases} \lambda x & \text{if } 0 \le x \le \lambda, \\ \frac{2\gamma\lambda x - x^2 - \lambda^2}{2(\gamma - 1)} & \text{if } \lambda < x < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } x \ge \gamma\lambda; \end{cases} \quad \mathrm{MCP}(x; \lambda, \gamma) \equiv \begin{cases} \lambda x - \frac{x^2}{2\gamma} & \text{if } 0 \le x \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & \text{if } x > \gamma\lambda. \end{cases}$$

We do not observe any clear relationship between features such as convexity, differentiability or the slope of h at 0, and performance of the estimator. Nonetheless, for many choices of rather simple functions h, our estimator provides a significant improvement over existing methods. In particular, most h functions that behave linearly for small x, namely $\log(1 + x)$ and x and their truncations, and additionally MCP and SCAD, always perform better than $x^{1.5}$ and x^2. This agrees with our discussion in Section 5.4.2, where $2 - a = 1$ is a reasonable choice of the power for small x; also see Section 7.3. However, we conclude that there is no real gain from making the function smoother by using MCP or SCAD.

Truncated Centered GGMs: For data from a truncated centered Gaussian distribution, we compare our generalized score matching estimator with various choices of h, to *SpaCE JAM* (SJ, Voorman et al., 2014), 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 *nonparanormal SKEPTIC* (Liu et al., 2012) with Kendall's τ. Recall that the choice of $h(x) = x^2$ corresponds to the estimator from Lin et al. (2016).

The ROC (*receiver operating characteristic*) curves for different estimators are shown in Figure 3 on Page 26. Each plotted curve corresponds to the average of 50 ROC curves, where the averaging is based on the vertical averaging from Algorithm 3 in Fawcett (2006), and is mean AUC-preserving. The x and y axes of each ROC curve represent the false positive and true positive rates at varying levels of penalty parameter λ, defined as

Centered, $n = 80$, multiplier 1.8647					
$\min(\log(1+x), c)$			$\min(x, c)$		
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					
$\min(\log(1+x), c)$			$\min(x, c)$		
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					
$\min(\log(1+x), c)$			$\min(x, c)$		
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).

$$\text{FPR} \equiv \frac{|\hat{S}_{\text{off}} \backslash S_{0,\text{off}}|}{m(m-1) - |S_{0,\text{off}}|} \qquad \text{and} \qquad \text{TPR} \equiv \frac{|\hat{S}_{\text{off}} \cap S_{0,\text{off}}|}{|S_{0,\text{off}}|},$$

where $S_{0,\text{off}} \equiv \{(i,j) : i \neq j \wedge \kappa_{0,ij} \neq 0\}$, and $\hat{S}_{\text{off}} \equiv \{(i,j) : i \neq j \wedge \hat{\kappa}_{ij} \neq 0\}$.

To reduce clutter, we only report the results for the top performing competing methods. In particular, results for nonparanormal SKEPTIC are omitted, as the method always performs the worst in our experiments. The corresponding means and standard deviations of AUCs (*areas under the curves*) over 50 curves are given in Table 1.

Looking at the mean AUCs, with the standard deviations in mind, all choices of h considered here 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 7.2.2.

Truncated Non-Centered GGMs: 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 7.1, and generate each component of $\boldsymbol{\mu}_0$ independently from the normal distribution with mean 0 and standard deviation 0.5.

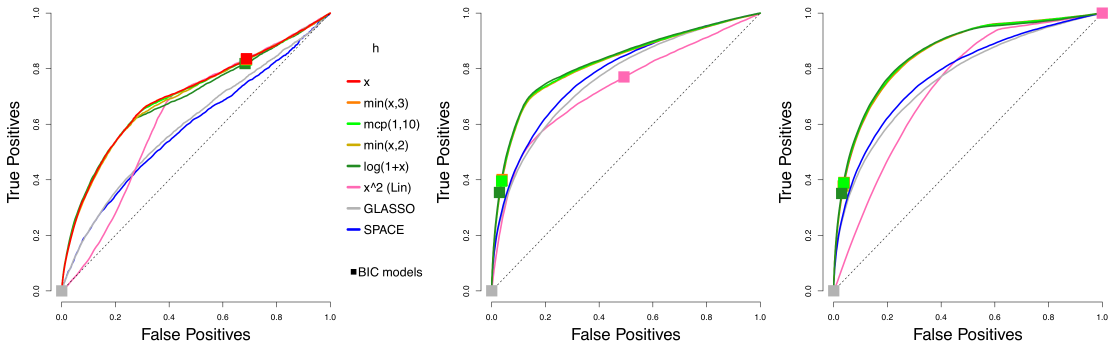
We compare the performance of our *profiled* estimator based on (19), with different h functions, but with no penalty on $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$, to SPACE, SpaCE JAM (SJ), GLASSO, and neighborhood selection (NS). As before, we consider 50 trials. Representative ROC curves are plotted in Figure 4, and the corresponding AUCs are summarized in Table 2.

Non-centered profiled, $n = 80$, multiplier 1.8647					
$\min(\log(1+x), c)$			$\min(x, c)$		
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					
$\min(\log(1+x), c)$			$\min(x, c)$		
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)		

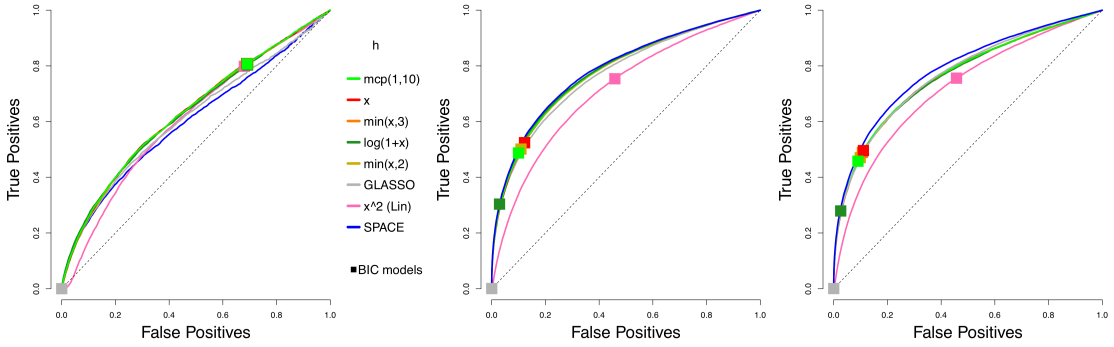
Non-centered profiled, $n = 1000$, multiplier 1.6438					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.764	0.018	∞	0.766	0.019
2	0.764	0.018	3	0.765	0.019
1	0.762	0.018	2	0.764	0.018
0.5	0.738	0.018	1	0.753	0.018
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.766	0.019	10	0.766	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).



(a) $n = 80$, mult $= 1.8647$ (b) $n = 1000$, mult $= 1$ (c) $n = 1000$, mult $= 1.6438$

Figure 3: Average ROC curves of our centered estimator with various choices of h, compared to SPACE and GLASSO, for the truncated centered GGM case; $m = 100$ variables and $n = 80$ or 1000 samples are considered. 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) $n = 80$, mult $= 1.8647$ (b) $n = 1000$, mult $= 1$ (c) $n = 1000$, mult $= 1.6438$

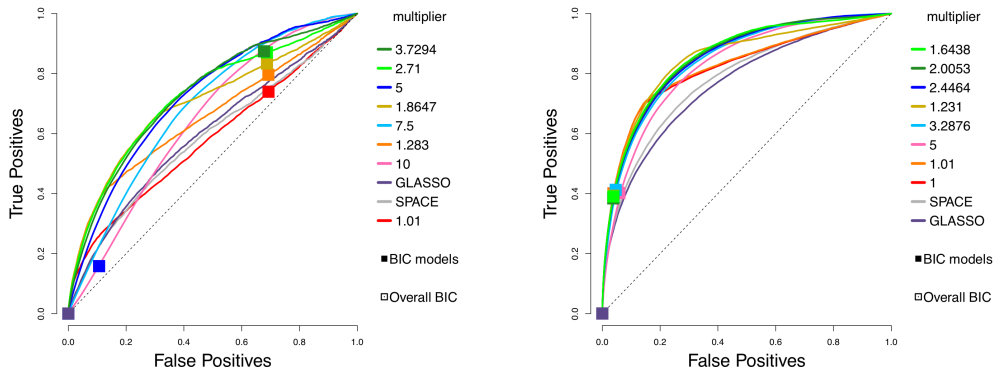
Figure 4: Average ROC curves of our non-centered profiled estimator with various choices of h, compared to SPACE and GLASSO, for the truncated non-centered GGM case; $m = 100$ variables and $n = 80$ or 1000 samples are considered. Squares indicate average true positive rate (TPR) and false positive rate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

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 $n = 80$. 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 deteriorate with a multiplier larger than 1; however, as we will see, there can be significant improvement in AUCs if we tune an additional parameter for the multiplier. 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 $h(x) = \min(x, 3)$.

7.2.2. Choice of multiplier

Truncated Centered GGMs: In Figure 5, the ROC curves for GLASSO, SPACE, and our estimator with $h(x) = \min(x, 3)$, but with different levels of amplification, via different choices of multipliers δ, are compared for the centered case of Section 7.2.1.

While Theorem 16 guarantees consistency only for $\delta < C(n, m)$, 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 $n = 1000$ (when $n = 1000$, $C(n, m)$ 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.



(a) $n = 80$ (b) $n = 1000$

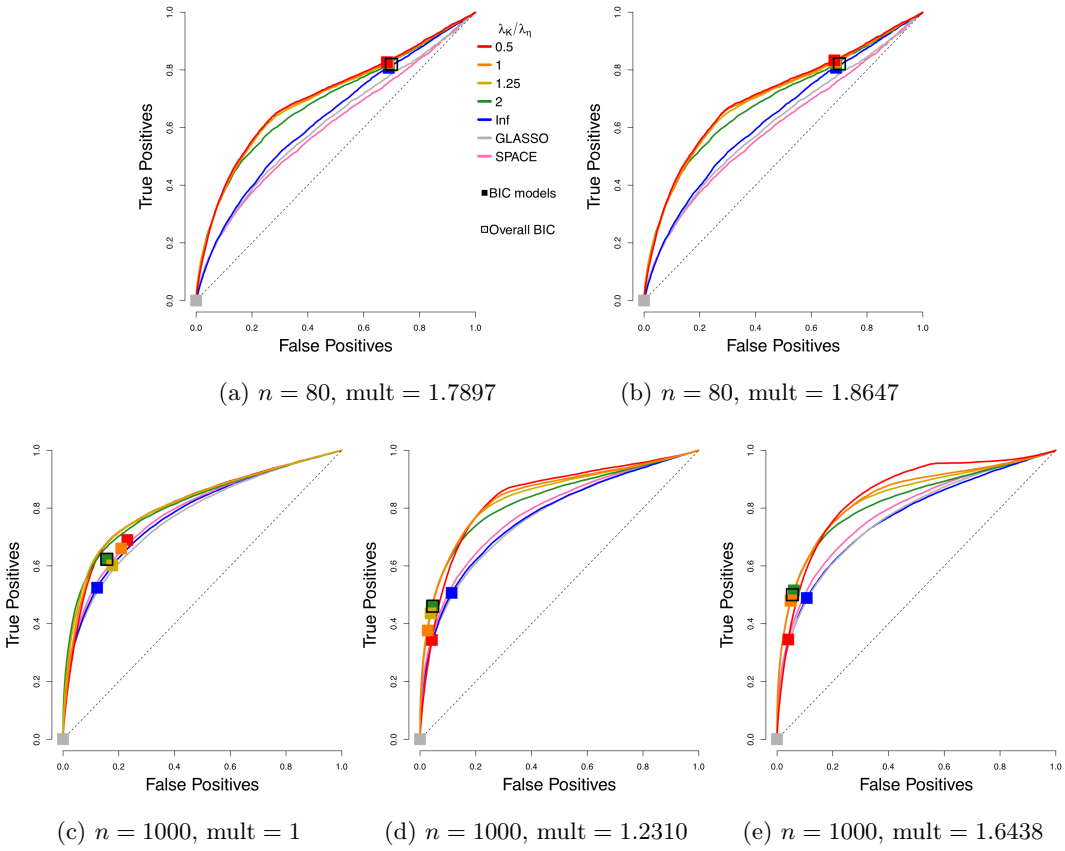
Figure 5: Performance of $\min(x, 3)$ for truncated centered GGMs using different multipliers, compared to GLASSO and SPACE, in the centered setting, $n = 80$ or 1000.

Truncated Non-Centered GGMs : In Figure 6, we consider the non-centered case of Section 7.2.1, 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 the choice of amplification in our estimator, we consider the upper-bound multiplier $C(n, m)$ from Theorem 17 as the default. We refer to this as *high* amplification. We also consider lower amplification, with $\delta = 2 - (1 + 24e \log m/n)^{-1}$, re-

ferred to as *medium*. 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 each run, the best model picked by BIC falls on the ROC curve, a few squares are off the curve in Figure 6 (c). This is because these squares correspond to the average of the true and false positive rates of the chosen BIC models over 50 runs, potentially due to multimodality of the distribution of the models. Nonetheless, 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 $= 1.7897$ (b) $n = 80$, mult $= 1.8647$

(c) $n = 1000$, mult $= 1$ (d) $n = 1000$, mult $= 1.2310$ (e) $n = 1000$, mult $= 1.6438$

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 the 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.

7.3. Other a/b Models

We now turn to the non-Gaussian ($a \neq 1$ or $b \neq 1$) setting. Based on the observations in Section 5.4.2, we focus on functions of type $\min(x^p, c)$ for some power $p > 0$ and truncation point $c > 0$. For simplicity, for the non-centered models we use the profiled estimator (19) (i.e., $\lambda_{\boldsymbol{\eta}} = 0$) and use the multiplier $C(n, m)$ in Theorem 16 for truncated GGMs as a guidance. We note that tuning over the $\lambda_{\boldsymbol{\eta}}$ parameter and the multiplier can potentially give a significant improvement as seen in Section 7.2.

These simulations suggest that among the class of functions of the form $\min(x^p, c)$, x^{2-a} or $\min(x^{2-a}, c)$ with a moderately large c can be used as the default choice of $h(x)$. This agrees with our findings in Section 7.2.1. We note that bounded h functions were only used in the proof for truncated GGMs, and picking a moderately large truncation point can correspond to having an untruncated power.

7.3.1. Exponential Setting

For the exponential models, $a = b = 1/2$. Since $a = b$, for both centered and non-centered settings, based on the principle in Section 5.4.2, choosing $h(x) = \min(x^{3/2}, c)$ satisfies (A1) and (A2) and also ensures that entries in $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are bounded (for small x), while choosing $h(x) = \min(\sqrt{x}, c)$ only guarantees (A1) and (A2).

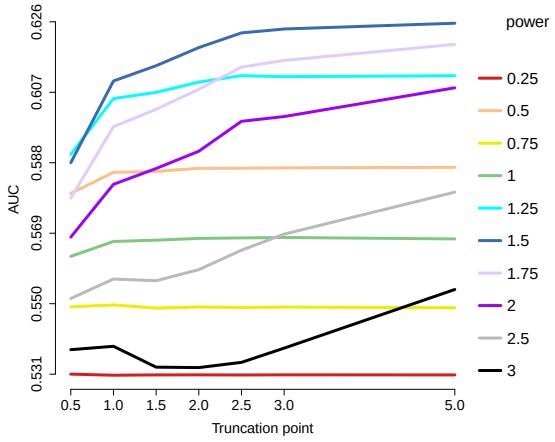
In Figure 7, we present the AUCs for the ROC curves of edge recovery with different choices of $h(x) = \min(x^{\text{pow}}, c)$. As before, we set $n = 80$ or 1000 and $m = 100$, but we use an $\boldsymbol{\eta}_0$ with each component uniformly equal to -0.5, 0 or 0.5; for $\boldsymbol{\eta}_0 \equiv \mathbf{0}$, we assume this information is known and use the centered estimator. The results suggest that $\text{pow} = 3/2 = 2 - a$ is the best choice of power. For this optimal choice, the performance improves with larger c, so x^{2-a} gives the best results. For sub-optimal powers, including truncation gives better results.

7.3.2. Gamma Setting

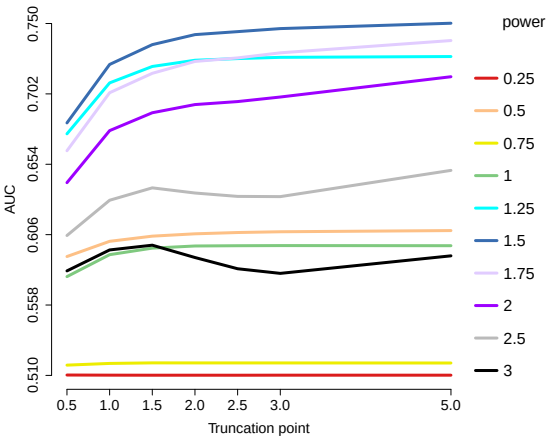
The centered gamma models reduce to the centered exponential models. Thus, in this section, we only consider the non-centered settings, with $a = 1/2$, $b = 0$. From Section 5.4.2, we have the following choices:

- $\min(x^2, c)$ both satisfies (A1)–(A2) and ensures $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are bounded;
- $\min(x^{\max\{3/2, 1-\min_j \eta_{0,j}\}}, c)$ ensures (A1)–(A2) and bounds $\boldsymbol{\Gamma}_{11}$ and $\boldsymbol{g}_1$; by default without prior information on $\boldsymbol{\eta}_0$ this is $\min(x^2, c)$;
- $\min(x^{3/2}, c)$ satisfies both conditions on the interaction part only ($\boldsymbol{x}^a$), but does not guarantee (A1)–(A2);
- $\min(x^{1/2}, c)$ satisfies the sufficient conditions for (A1)–(A2) on the interaction only.

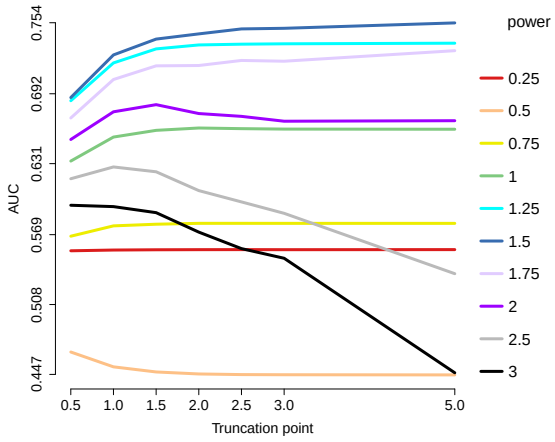
The results are shown in Figure 8, where we consider $n = 80$, 1000, and $\boldsymbol{\eta} = \pm 0.5\mathbf{1}_{100}$. They suggest that $\text{pow} = 2 - a = 1.5$ works consistently well, although slightly outperformed by 1 and 1.25 in one case. As in the exponential case, with the optimal power it is beneficial to choose a large truncation point, or work with an untruncated power $x^{1.5}$. We conclude that the performance is likely only dependent on the $(2 - a)$ power requirement for the ${\boldsymbol{x}^a}^\top \mathbf{K} \boldsymbol{x}^a$ part or $2 - \min_j \eta_{0,j}$; simulations in the next section rule out the possibility of the latter.



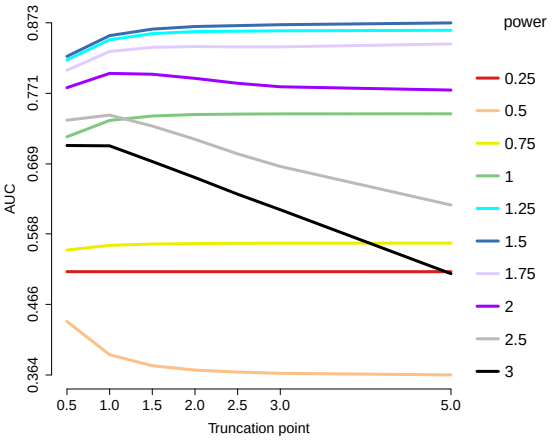
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



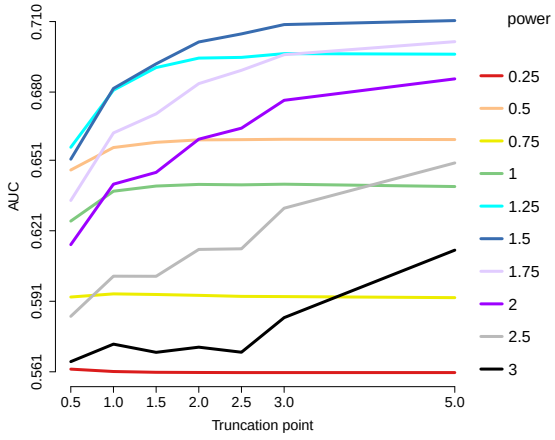
(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



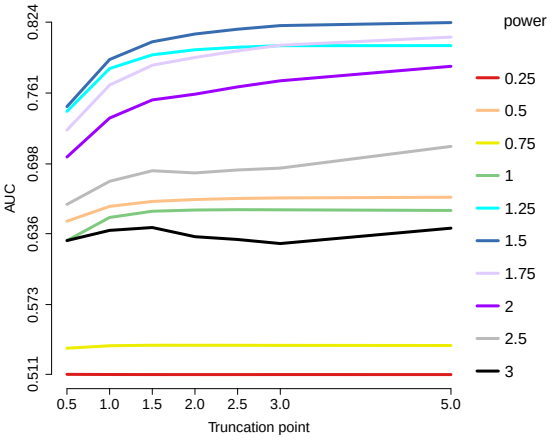
(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator



(d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator



(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator



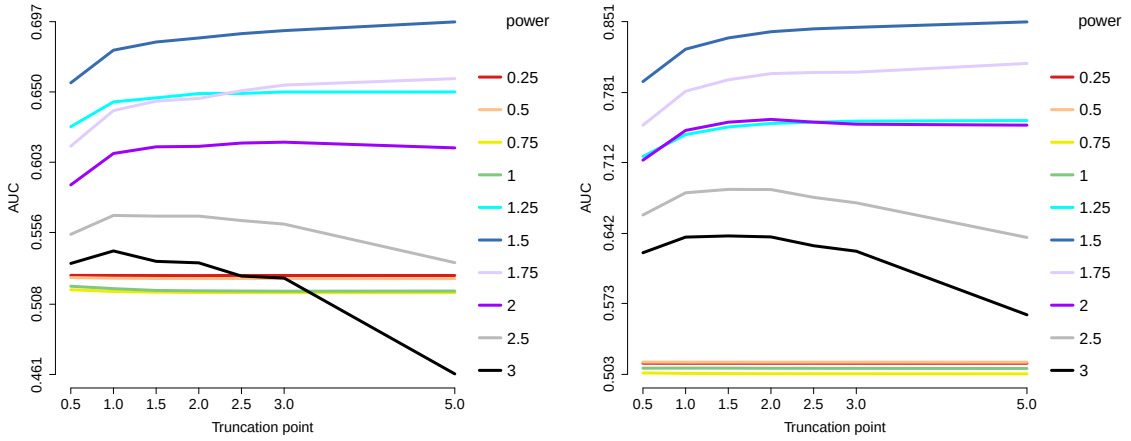
(f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure 7: AUCs for edge recovery using generalized score matching for the exponential models. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.

7.3.3. Other Choices of a and b

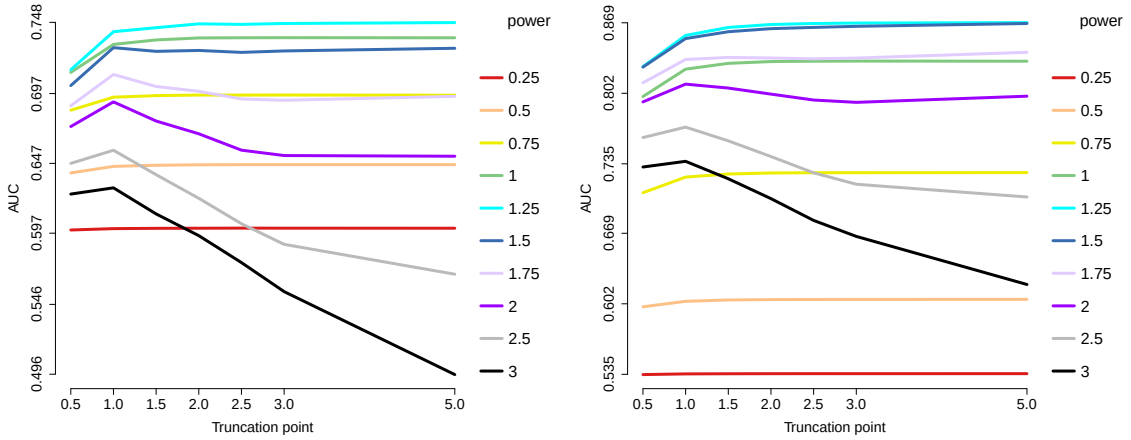
In this section, we consider other choices of a and b. Specifically, $a = 3/2$ and $b = 1/2$ or 0. These combinations are chosen to confirm, in a more extreme setting, that the performance is mainly determined by requirements on the power based on a, which correspond to choosing a power of $1 - a$ or $2 - a$, but not those on b (or on $\boldsymbol{\eta}$ when $b = 0$) that correspond to $1 - b$ and $2 - b$. The relationship between these two settings is analogous to that between the exponential and gamma models (same a, b nonzero/zero).

The results are shown in Figures 9 and 10, and indeed confirm that $x^{2-a} = x^{0.5}$ consistently gives the optimal results, even though $\boldsymbol{\eta}^\top \boldsymbol{x}^b$ is in favor of $x^{2-b} = x^{1.5}$ for $b = 0.5$, and $\boldsymbol{\eta}^\top \log(\boldsymbol{x})$ is in favor of x^2 or at least $x^{1-\min_j \eta_{0,j}}$ when $b = 0$. There are two possible explanations for the optimality of $2 - a$ over $\max\{2 - a, 2 - b\}$ or $\max\{2 - a, 1 - \min_j \eta_{0,j}\}$: (1) The AUC metric is measured only on our interest, edge recovery for the interaction matrix, which only depends on $\boldsymbol{x}^a$; (2) using the profiled estimator weakens the effect of b.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

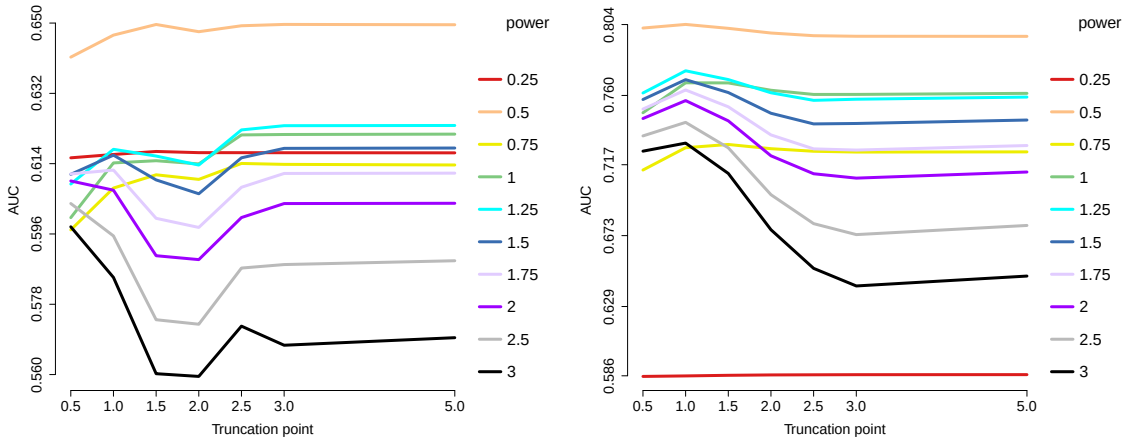
(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

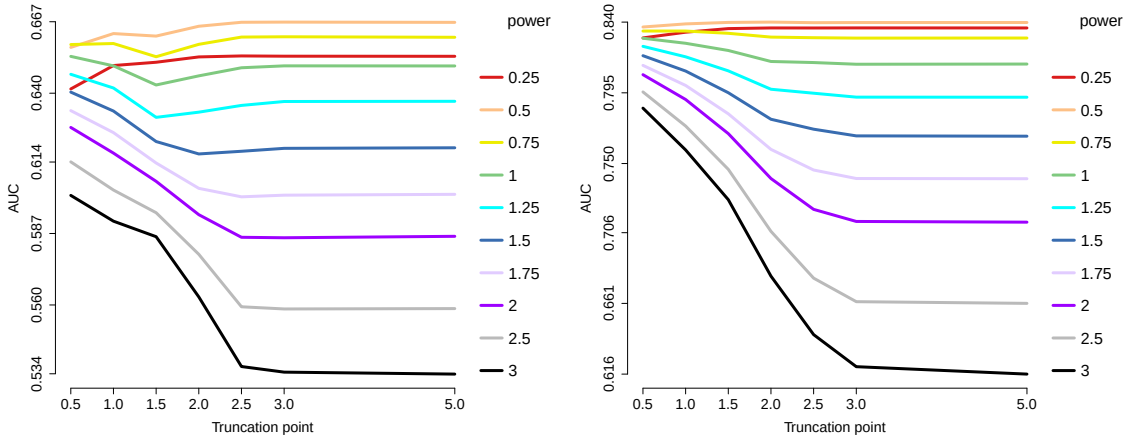
(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure 8: AUCs for edge recovery using generalized score matching for the gamma models. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.



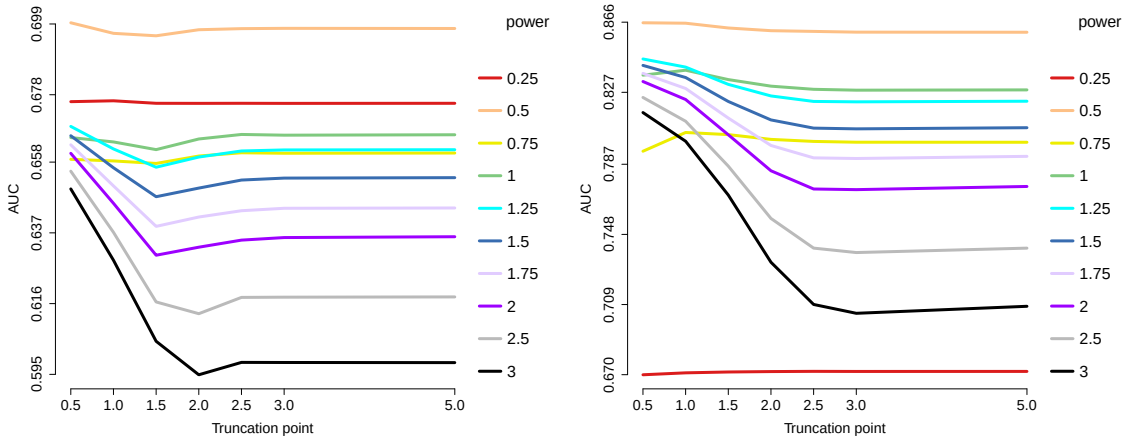
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator

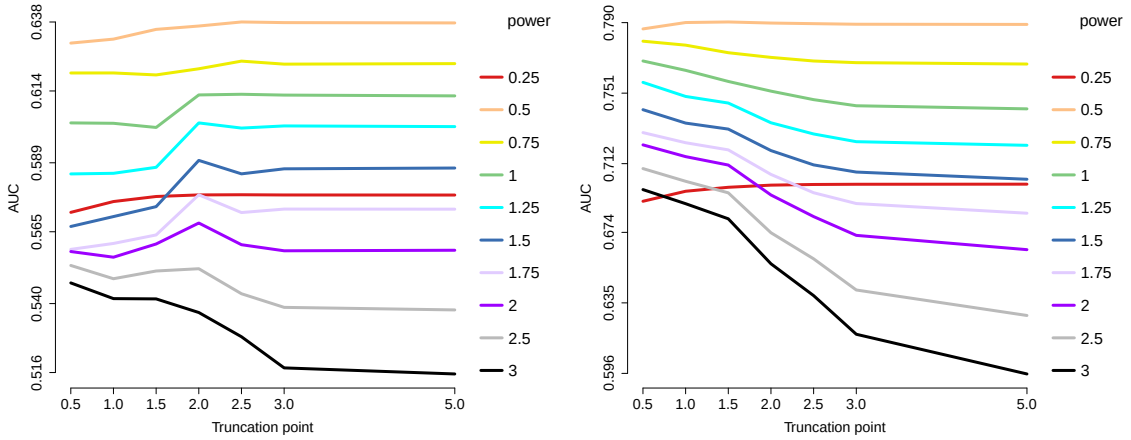
(d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator



(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

(f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure 9: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 1/2$. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

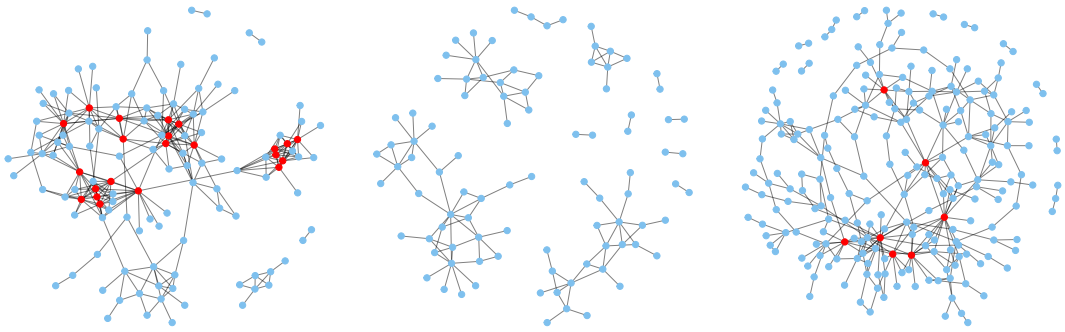
Figure 10: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 0$. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.

7.4. 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), since the same model is considered therein. The data consists of $n = 487$ prostate adenocarcinoma samples from The Cancer Genome Atlas (TCGA) data set. 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 7.2.2. 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).

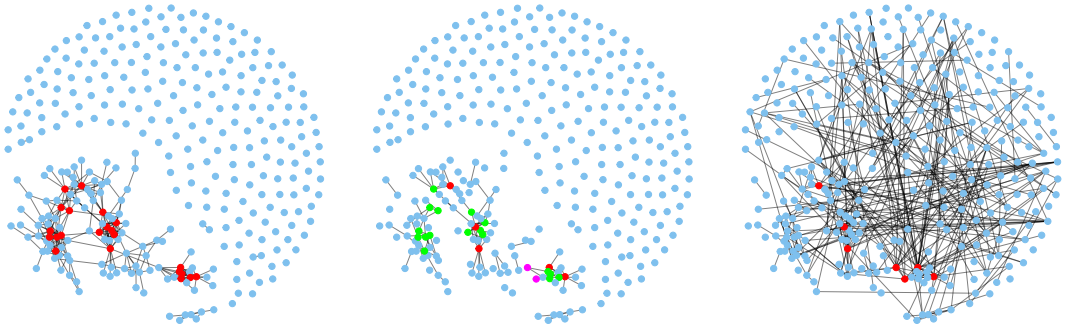
We compare our graph to the one in Lin et al. (2016), which corresponds to $h(x) = x^2$ with no multiplier. Shown in Figure 11 are the estimated graphs, with their intersection in the middle. To improve visualization, isolated nodes are removed and the layouts are optimized for each plot. Red-colored points are the "hub nodes", namely nodes with degree at least 10. In Figure 12, 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 that 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, so $P(R = r) = \frac{\binom{m}{r}\binom{m(m-1)/2-m}{m-r}}{\binom{m(m-1)/2}{m}}$. For $m = 333$ the probability of at least 117 common edges is essentially zero. 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.



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

Figure 11: 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 12: 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, and the magenta ones are hub nodes in graph (c) only.

The graph using $h(x) = \min(x, 3)$ has much more isolated nodes (204) than the other (108), and has a slightly smaller max degree (16 versus 19). Table 3 provides another way of comparing between the two graphs by listing the genes with the highest node degrees.

$\min(x, 3)$ 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). As noted in Lin et al. (2016), genes with high node-degrees are known to be important in biological

networks (Carter et al., 2004; Jeong et al., 2001; Han et al., 2004). 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 activity of MMP-2 can be useful in diagnosis, therapy, and assessment of malignant progression in prostate cancer.

- 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).

We note that the two methods indeed use different estimators (different h functions and multipliers), and it is thus not surprising to see that some of the top genes by one method are not among those for the other. In particular, CCNE2, BRCA2, SKP2 and STAT5B, while previously reported as newly discovered in Lin et al. (2016), are dropped by our new analysis. Testing and inference (potentially using bootstrapping) is an important problem but is beyond the scope of this paper.

8. Discussion

In this paper, we proposed a generalization of the score matching estimator of Hyvärinen (2007), based on scaling the log-gradients to be matched with a suitably chosen function $\boldsymbol{h}$. The generalization retains the advantages of Hyvärinen's method: Estimates can be computed without knowledge of normalizing constants, and for canonical parameters of exponential families, the estimation loss is a quadratic function.

For high-dimensional exponential family graphical models, following Lin et al. (2016), we add an ℓ_1 penalty to regularize the generalized score matching loss. One practical issue that is overlooked in Lin et al. (2016) is the fact that the score matching loss can be unbounded below for a small tuning parameter, when the dimension m exceeds the sample size 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 we give an upper bound on that multiplier that guarantees consistency.

As examples we consider *pairwise interaction power models* on the non-negative orthant $\mathbb{R}^m_+$. Specifically, the considered models are exponential families in which the log density is

the sum of pairwise interactions between entries in of powers $\boldsymbol{x}^a$ plus linearly weighted effects $\boldsymbol{x}^b$, or $\log(\boldsymbol{x})$ when $b = 0$. Our main interest is in the matrix of interaction parameters whose support determines the distributions' conditional independence graph. The considered framework covers truncated normal distributions ($a = b = 1$), exponential square root graphical models ($a = b = 1/2$) from Inouye et al. (2016), as well as a class of multivariate gamma distributions ($a = 1/2$, $b = 0$).

In the case of multivariate truncated normal distributions, where the conditional independence graph is given by the underlying Gaussian inverse covariance matrix, the sample size required for the consistency of our method using bounded $\boldsymbol{h}$ is $\Omega(d^2 \log m)$, where d is the degree of the graph. This matches the rates for Gaussian graphical models in Ravikumar et al. (2011) and Lin et al. (2016). In contrast, the sample complexity for truncated Gaussian models given in Lin et al. (2016) is $\Omega(d^2 \log^8 m)$.

For the considered class of pairwise interaction models, we recommend using the function $\boldsymbol{h}$ with coordinates $h_j(x) = \min\left(x^{2-a}, c\right)$ for some moderately large c, or simply $h_j(x) = x^{2-a}$. While this choice is effective, it would be an interesting problem for future work to develop a method that adaptively chooses an optimized function $\boldsymbol{h}$ from data.

Acknowledgments

This work was partially supported by grant DMS/NIGMS-1561814 from the National Science Foundation (NSF). AS also gratefully acknowledges funding by grant R01-GM114029 from the National Institute of Health (NIH).

Appendix A. Proofs

A.1. Proof of Theorem 3

The following integration by parts lemma is used in the proof of Theorem 3.

Lemma 19 *Let $f, g : \mathbb{R}_+ \to \mathbb{R}$ be functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$. Then*

$$\lim_{x \nearrow +\infty} f(x)g(x) - \lim_{x \searrow 0^+} f(x)g(x) = \int_0^\infty f(\boldsymbol{x}) \frac{\mathrm{d}g(x)}{\mathrm{d}x}\,\mathrm{d}x + \int_0^\infty g(\boldsymbol{x}) \frac{\mathrm{d}f(x)}{\mathrm{d}x}\,\mathrm{d}x.$$

Proof This is an analog of Lemma 4 from Hyvärinen (2005) that can be proved by integrating the partial derivatives, and follows from the fundamental theorem of calculus for absolutely continuous functions and the product rule. In particular, we work on integrals in bounded $[0, c]$, where the product of two absolutely continuous functions in a bounded interval is again absolutely continuous, and the result is then obtained by letting $c \nearrow +\infty$. ■

Proof [Proof of Theorem 3] Recall the following assumptions from Section 2.2:

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

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

Without explicitly writing the domains $\mathbb{R}_+$ or $\mathbb{R}_+^m$ in all integrals, by (6) we have

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

where A will simply appear in the final display as is, C is a constant as it only involves the true pdf p_0, and we wish to simplify B by integration by parts. We can split the integral into these three parts since A and C are assumed finite in the first part of (A2), and the integrand in B is integrable since $|2ab| \le a^2 + b^2$. Thus, by linearity and Fubini's theorem, we can write

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

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

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

But, we assume p_0 and p are twice continuously differentiable, for every $j = 1, \ldots, m$ and fixed $\boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}$. Hence, in every bounded sub-interval of $\mathbb{R}_+$, $p_0(\boldsymbol{x}_{-j}; x_j)$ is an absolutely continuous function of x_j, $\partial_j \log p(\boldsymbol{x}_{-j}, x_j) = \partial_j p(\boldsymbol{x}_{-j}, x_j)/p(\boldsymbol{x}_{-j}, x_j)$ is a continuously differentiable (and hence absolutely continuous) function of x_j by the quotient rule. Thus $h_j(x_j)\partial_j \log p(\boldsymbol{x}_{-j}; x_j)$ is also absolutely continuous by the absolute continuity assumption on h_j. Then, by Lemma 19, where we take $f \equiv p_0(\boldsymbol{x}_{-j}; x_j)$ and $g \equiv h_j(x_j)\partial_j \log p(\boldsymbol{x}_{-j}; x_j)$ as functions of x_j, followed by assumption (A1),

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

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 is a constant that does not depend on p. ■

A.2. Proof of Theorems and Examples in Section 3

Proof [Proof of Theorem 5] For exponential families and under the assumptions, the empirical loss $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ in (8) becomes (up to an additive constant)

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

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

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top, \tag{34}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\right]. \tag{35}$$

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

Proof [Proof of Theorem 6] By Theorem 5, $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}\boldsymbol{\theta} - \boldsymbol{g}^\top\boldsymbol{\theta} + \text{const}$. The minimizer of $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ is thus available in the unique closed form $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$ as long as $\boldsymbol{\Gamma}$ is invertible (C1). Since $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are sample averages, the weak law of large numbers yields that $\boldsymbol{\Gamma} \to_p \mathbb{E}_{p_0}\boldsymbol{\Gamma} \equiv \boldsymbol{\Gamma}_0$ and $\boldsymbol{g} \to_p \mathbb{E}_{p_0}\boldsymbol{g} \equiv \boldsymbol{g}_0$, where existence of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{g}_0$ is assumed in (C2). Since $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \mathbb{E}[\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})] = \mathbb{E}[\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top\boldsymbol{\theta}] = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}_0\boldsymbol{\theta} - \boldsymbol{g}_0\boldsymbol{\theta}$ and we know $\boldsymbol{\theta}_0$ minimizes $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ by definition, by the first-order condition we must have $\boldsymbol{\Gamma}_0\boldsymbol{\theta}_0 = \boldsymbol{g}_0$. Then by the Lindeberg-Lévy central limit theorem,

$$\sqrt{n}(\boldsymbol{g}(\mathbf{x}) - \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0) \to_d \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_0),$$

where $\boldsymbol{\Sigma}_0 \equiv \mathbb{E}_{p_0}[(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))^\top]$, as long as $\boldsymbol{\Sigma}_0$ exists (C2). Thus, by Slutsky's theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \equiv \sqrt{n}(\boldsymbol{\Gamma}(\mathbf{x})^{-1}(\boldsymbol{g}(\mathbf{x}) - \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0)) \to_d \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0^{-1}),$$

as long as $\mathbf{\Gamma}_0$ is invertible (C2).

For the second half of the theorem, (C2) $\mathbb{E}_{p_0}\mathbf{\Gamma}(\mathbf{x}) < \infty$ and $\mathbb{E}_{p_0}\boldsymbol{g}(\mathbf{x}) < \infty$ implies $\mathbb{E}_{p_0}|\mathbf{\Gamma}(\mathbf{x})| < \infty$ and $\mathbb{E}_{p_0}|\boldsymbol{g}(\mathbf{x})| < \infty$, so by strong law of large numbers (and a union bound on at most k^2 null sets)

$$\mathbf{\Gamma}(\mathbf{x}) \to_{\text{a.s.}} \mathbf{\Gamma}_0, \quad \boldsymbol{g}(\mathbf{x}) \to_{\text{a.s.}} \boldsymbol{g}_0.$$

Then outside a null set,

$$\hat{\boldsymbol{\theta}} \equiv \mathbf{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x}) \to_{\text{a.s.}} \mathbf{\Gamma}_0^{-1}\boldsymbol{g}_0 = \boldsymbol{\theta}_0.$$

■

Proof [Proof for Example 3.1] We choose to estimate $\theta \equiv \mu/\sigma^2$. Then by (11) and (12),

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

By Theorem 6,

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

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

$$\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X-\mu_0)] \\
=& \int_0^\infty h'(x)h(x)(x-\mu_0)p(x)\,\mathrm{d}x \\
=& \int_0^\infty h(x)(x-\mu_0)p(x)\,\mathrm{d}h(x) \\
=& \ h^2(x)(x-\mu_0)p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu_0)p(x) \\
=& -\int h^2(x)p(x)\,\mathrm{d}x - \int h(x)h'(x)(x-\mu_0)p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu_0)^2}{\sigma^2}p(x)\,\mathrm{d}x,
\end{aligned}$$

where the last step follows from the assumptions $\lim_{x\searrow 0^+} h(x) = 0$ and $\lim_{x\nearrow+\infty} h^2(x)(x-\mu_0)p_{\mu_0}(x) = 0$. So

$$\mathbb{E}_0[h(X)h'(X)(X-\mu_0)] = \frac{\mathbb{E}[h^2(X)((X-\mu_0)^2/\sigma^2 - 1)]}{2}. \tag{36}$$

The asymptotic variance is thus

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

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

Proof [Proof for Example 3.2] We estimate $\theta \equiv 1/\sigma^2$. By (11) and (12),

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

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

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

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

$$\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X-\mu)^3] \\
=\ &\int_0^\infty h'(x)h(x)(x-\mu)^3 p(x)\,\mathrm{d}x \\
=\ &\int_0^\infty h(x)(x-\mu)^3 p(x)\,\mathrm{d}h(x) \\
=\ &h^2(x)(x-\mu)^3 p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu)^3 p(x) \\
=\ &-\int h(x)h'(x)(x-\mu)^3 p(x)\,\mathrm{d}x - 3\int h^2(x)(x-\mu)^2 p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu)^4}{\sigma_0^2} p(x)\,\mathrm{d}x,
\end{aligned}$$

where the last step follows from the assumptions $\lim_{x \searrow 0^+} h(x) = 0$ and $\lim_{x \nearrow +\infty} h^2(x)(x-\mu)^3 p_{\sigma_0^2}(x) = 0$. Combining this with (36) 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. ■

A.3. Proof of Theorems in Section 5

Proof [Proof of Theorem 9]
Case $b \neq 0$: We use a strategy similar to that of Inouye et al. (2016). Let $\mathcal{V}_1 = \{\boldsymbol{v} : \|\boldsymbol{v}\|_1 = 1, \boldsymbol{v} \in \mathbb{R}_+^m\}$. Then by Fubini-Tonelli the normalizing constant is,

$$\begin{aligned}
&\int_{\mathbb{R}_+^m} \exp\left(\boldsymbol{\eta}^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b} - \frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a\right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{V}_1}\int_0^\infty \exp\left(\boldsymbol{\eta}^\top \frac{z^b\boldsymbol{v}^b - \mathbf{1}_m}{b} - \frac{1}{2a}z^{2a}\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a\right) \mathrm{d}z\,\mathrm{d}\boldsymbol{v} \\
&\propto \int_{\mathcal{V}_1}\int_0^\infty \exp\left(z^b(\boldsymbol{\eta}^\top\boldsymbol{v}^b)/b - z^{2a}(\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a)/(2a)\right) \mathrm{d}z\,\mathrm{d}\boldsymbol{v}.
\end{aligned}$$

Here $\mathcal{V}_1$ is compact and the inner integral, if finite, is continuous in $\boldsymbol{v}$. It thus suffices to show that the inner integral is finite at every single $\boldsymbol{v} \in \mathcal{V}_1$.

Fixing $\boldsymbol{v} \in \mathcal{V}_1$, write $A \equiv A(\boldsymbol{v}) \equiv \boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a/(2a)$ and $B \equiv B(\boldsymbol{v}) \equiv (\boldsymbol{\eta}^\top\boldsymbol{v}^b)/b$. We need to show that

$$N(A, B, a, b) \equiv \int_0^\infty \exp\left(-Az^{2a} + Bz^b\right) \mathrm{d}z < +\infty.$$

Recall that (CC1) $\boldsymbol{v}^\top\mathbf{K}\boldsymbol{v} > 0$ for all $\boldsymbol{v} \in \mathbb{R}_+^m \backslash \{\mathbf{0}\}$, so $A > 0$.

(i) Suppose $B \leq 0$. Then $N(A, B, a, b) \leq \int_0^\infty \exp(-Az^{2a})\,\mathrm{d}z = A^{-a/2}\Gamma(1 + 1/(2a))$, a finite constant since $A > 0$ and $a > 0$.

(ii) Suppose $B > 0$. We first want to bound $\exp(-Az^{2a} + Bz^b) \leq N_0 \exp(-Az^{2a}/2)$ by some finite constant $N_0 > 0$, so that $N(A, B, a, b) \leq N_0 \int_0^\infty \exp(-Az^{2a}/2)\,\mathrm{d}z$, a finite constant for $a > 0$. Thus, it remains to give conditions so that $\exp(-Az^{2a}/2 + Bz^b)$ is bounded by some finite constant N_0, which by continuity only requires a finite limit as $z \searrow 0$ and as $z \nearrow +\infty$. As $z \nearrow +\infty$, $Bz^b \nearrow +\infty$, while $-Az^{2a}/2 \searrow -\infty$. We thus need $b < 2a$ so that the sum of the two does not go to positive infinity. On the other hand, as $z \searrow 0$, $-Az^{2a}/2 \nearrow 0$, so we need $b > 0$, otherwise $z^b \nearrow +\infty$. In conclusion, we require that $2a > b > 0$.

It thus suffices to require (CC1) and (CC2) $2a > b > 0$ to eliminate restrictions on B, and hence on $\boldsymbol{\eta}$. That is, $\boldsymbol{\eta}$ can take value in the entirety of $\mathbb{R}^m$.

Case $b = 0$: Again in (CC1) we assume $\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} > 0$ for all $\boldsymbol{v} \in \mathbb{R}^m_+\backslash\{\mathbf{0}\}$. Since $\mathcal{V}_2 \equiv \{\boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1, \boldsymbol{v} \in \mathbb{R}^m_+\}$ is compact and $\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v}$ is continuous in $\boldsymbol{v}$ and strictly positive on $\mathcal{V}_2$, the image of $\mathcal{V}_2$ under $\boldsymbol{v}^\top \mathbf{K}\boldsymbol{v}$ is a compact subset of $(0, \infty)$, i.e. $N_{\mathbf{K}} \equiv \min_{\boldsymbol{v}\in\mathbb{R}^m_+\backslash\{\mathbf{0}\}} \boldsymbol{v}^\top \mathbf{K}\boldsymbol{v}/\boldsymbol{v}^\top\boldsymbol{v} \equiv \min_{\boldsymbol{v}\in\mathcal{V}_2} \boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} > 0$. We thus have

$$
\begin{aligned}
&\int_{\mathbb{R}^m_+} \exp\left(\boldsymbol{\eta}^\top \log(\boldsymbol{x}) - \frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a\right) \mathrm{d}\boldsymbol{x} \\
\leq &\int_{\mathbb{R}^m_+} \exp\left(\boldsymbol{\eta}^\top \log(\boldsymbol{x}) - \frac{N_{\mathbf{K}}}{2a}\boldsymbol{x}^{a\top}\boldsymbol{x}^a\right) \mathrm{d}\boldsymbol{x} \\
= &\prod_{j=1}^m \int_0^\infty \exp\left(\eta_j \log(x_j) - \frac{N_{\mathbf{K}}}{2a} x_j^{2a}\right) \mathrm{d}x_j \\
= &\prod_{j=1}^m \left[\Gamma\left(\frac{\eta_j + 1}{2a}\right) \frac{(N_{\mathbf{K}}/2a)^{-\frac{\eta_j+1}{2a}}}{2a}\right],
\end{aligned}
$$

where the integration follows by change of variable and requires $a > 0$. Assuming $a > 0$, the last quantity is finite if and only if $\boldsymbol{\eta} \succ -\mathbf{1}_m$, by definition of the gamma function.

In conclusion, given conditions (CC1) $\min_{\boldsymbol{v}\in\mathbb{R}^m_+\backslash\{\mathbf{0}\}} \boldsymbol{v}^\top \mathbf{K}\boldsymbol{v} > 0$, (CC2) $2a > b > 0$, and (CC3) $a > 0$, $b = 0$ and $\boldsymbol{\eta} \succ -\mathbf{1}_m$, the unnormalized density (16) has a finite normalizing constant when (CC1) and (CC2) both hold, or (CC1) and (CC3) both hold.

The centered settings, where the term involving $\boldsymbol{x}^b$ is excluded, can be considered as a special case of both (1) and (2) with $\boldsymbol{\eta} \equiv \mathbf{0}$, and thus (CC1) and $a > 0$ are sufficient. ■

Proof [Proof of Theorem 11] Recall assumptions (A1) and (A2):

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

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

Let $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ be the true parameters so that $p_0 \in \mathcal{P}_+$, with $\mathcal{P}_+$ corresponding to a parameter space in which all parameters satisfy the conditions for a finite normalizing constant. We now give sufficient conditions for h to satisfy (A1) and (A2).

Conditions for (A1): Fix $j = 1, \ldots, m$ and $\boldsymbol{x}_{-j} \in \mathbb{R}^{m-1}_+$. We show that the conditions on h_j imply that the limits go to 0 as $x_j \nearrow +\infty$ and as $x_j \searrow 0^+$, which is stronger than (A1); in fact, from (37) below, the limits cannot go to a nonzero finite constant assuming an h with polynomial tail, since $a > 0$ and $B_1 \equiv \kappa_{0,jj} > 0$ for all j. Now,

$$
\begin{aligned}
&p_0(\boldsymbol{x})h_j(x_j)\partial_j \log p(\boldsymbol{x}) \\
\propto\ &h_j(x_j) \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}_0\boldsymbol{x}^a + \boldsymbol{\eta}_0^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b}\right) \partial_j \left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a + \boldsymbol{\eta}^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b}\right)
\end{aligned}
$$

$$\propto h_j(x_j)\exp\left(-\frac{1}{a}(\boldsymbol{k}_{0,j,-j}^\top \boldsymbol{x}_{-j}^a)x_j^a - \frac{\kappa_{0,jj}}{2a}x_j^{2a} + \eta_{0,j}\frac{x_j^b-1}{b}\right)\times$$
$$\left(-\boldsymbol{k}_{j,-j}^\top \boldsymbol{x}_{-j}^a x_j^{a-1} - \kappa_{jj}x_j^{2a-1} + \eta_j x_j^{b-1}\right)$$
$$\equiv h_j(x_j)\exp\left(\frac{A_1 x_j^a}{a} + \frac{B_1 x_j^{2a}}{2a} + C_1\frac{x_j^b-1}{b}\right)\left(A_2 x_j^{a-1} + B_2 x_j^{2a-1} + C_2 x_j^{b-1}\right), \tag{37}$$

where $A_1 \equiv -\boldsymbol{k}_{0,j,-j}^\top \boldsymbol{x}_{-j}^a$, $A_2 \equiv -\boldsymbol{k}_{j,-j}^\top \boldsymbol{x}_{-j}^a$, $B_1 \equiv -\kappa_{0,jj} < 0$ and $B_2 \equiv -\kappa_{jj} < 0$ by condition (CC1). Finally $C_1 \equiv \eta_{0,j}$, $C_2 \equiv \eta_j$.

(1) Let $x_j \nearrow +\infty$. If $b > 0$, since $2a > b > 0$ and $B_1 < 0$, the exponential term in (37) decreases to 0 exponentially and its reciprocal dominates any polynomial functions. Thus, the entire product goes to 0 if $h_j(x_j)$ grows no faster than polynomially as $x_j \nearrow +\infty$. If $b = 0$, the $C_1 \log x_j$ term is again dominated by $B_1 x_j^{2a}/(2a)$, and the same conclusion holds.

(2) Let $x_j \searrow 0$.

(i) Let $b > 0$. Then the exponential term in (37) goes to constant $\exp(-C_1/b)$, and we only need

$$\lim_{x_j \searrow 0^+} h_j(x_j)(A_2 x_j^{a-1} + B_2 x_j^{2a-1} + C_2 x_j^{b-1}) = 0. \tag{38}$$

- If $a > 1$ and $b > 1$, the second term in (38) is a polynomial with three terms having powers $\geq \min\{a-1, b-1\}$. The product goes to zero if and only if $h_j(x_j) = o(x_j^{\max\{1-a,1-b\}})$ as $x_j \searrow 0$. Note that this is satisfied by any h_j that has a finite right limit at 0.
- If $a = 1$ and $b \geq 1$, or $a \geq 1$ and $b = 1$, then the second term in (38) is a polynomial of non-negative power plus a potentially nonzero constant. A sufficient condition for (38) is thus $\lim_{x_j \searrow 0} h_j(x_j) = 0$.
- If $a < 1$ or $b < 1$, then the second part in (38) is a polynomial having terms with negative degree $\geq \min\{a-1, b-1\}$. To counteract this a sufficient condition is $h_j(x_j) = o(x_j^{\max\{1-a,1-b\}})$.

In conclusion, $\lim_{x_j \searrow 0^+} p_0(\boldsymbol{x})h_j(x_j)\partial_j \log p(\boldsymbol{x}) = 0$ if and only if

$$\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{\max\{1-a,1-b\}} = 0.$$

(ii) Now assume $b = 0$. Then, (37) now becomes

$$h_j(x_j)\exp\left(A_1 x_j^a/a + B_1 x_j^{2a}/(2a) + C_1 \log x_j\right)\left(A_2 x_j^{a-1} + B_2 x_j^{2a-1} + C_2/x_j\right).$$

With $C_1 \log x_j$ dominating, the exponential part scales as $x_j^{C_1}$. We thus require

$$\lim_{x_j \searrow 0^+} h_j(x_j)(A_2 x_j^{a-1+C_1} + B_2 x_j^{2a-1+C_1} + C_2 x_j^{C_1-1}) = 0,$$

which by the previous discussion on (38) holds if and only if

$$\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{1-C_1} = 0$$

since $1 - a - C_1 < 1 - C_1$.

In summary, (A1) is satisfied if $h_j(x_j)$ grows at most polynomially as $x_j \nearrow +\infty$, and $\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{\max\{1-a,1-b\}} = 0$ if $b > 0$, or $\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{1-\eta_{0,j}} = 0$ if $b = 0$.

Conditions for (A2): For (A2), we consider powers of x as the h functions for simplicity; conclusions for other functions that have the same tail behavior (big-O scaling) as $x \searrow 0$ and $x \nearrow +\infty$ follow similarly. Sufficiency results for piecewise power functions follow by partitioning, and similarly for other functions h whose function values and derivatives can be bounded by those of some piecewise power function (e.g. truncated powers), since (A2) is on integrability of products involving positive powers of h and h'.

Let $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ be the true parameters from the parameter space that satisfies the conditions for finite normalizing constant. By part (2) of the proof of Theorem 9, the assumption that $\mathbf{K}_0$ satisfies (CC1) implies that $\min_{\boldsymbol{v} \in \mathbb{R}_+^m \setminus \{\mathbf{0}\}} \boldsymbol{v}^\top \mathbf{K}_0 \boldsymbol{v} / \boldsymbol{v}^\top \boldsymbol{v} \equiv N_{\mathbf{K}_0} > 0$. Then we have the following decomposition

$$\begin{aligned} p_{\mathbf{K}_0,\boldsymbol{\eta}_0}(\boldsymbol{x}) &\equiv \exp\left(-\frac{1}{2a} \boldsymbol{x}^{a\top} \mathbf{K}_0 \boldsymbol{x}^a + \boldsymbol{\eta}_0^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b} \right) \\ &\leq \prod_{j=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b} \right). \end{aligned}$$

Then for any other $\mathbf{K}$ and $\boldsymbol{\eta}$ in the parameter space, for the first part of (A2) it suffices to show for any $j = 1, \ldots, m$ that $D < \infty$, where

$$\begin{aligned} D \equiv \int_{\mathbb{R}_+^m} \prod_{j=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b} \right) h_j(x_j) \times \\ \left(-\kappa_{jj} x_j^{2a-1} - \sum_{i \neq j} \kappa_{ji} x_i^a x_j^{a-1} + \eta_j x_j^{b-1} \right)^2 \mathrm{d}\boldsymbol{x} \\ \geq \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) h_j(x_j) \left(\partial_j \log p(\boldsymbol{x}) \right)^2 \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Note that

$$\begin{aligned} &\left(-\kappa_{jj} x_j^{2a-1} - \sum_{i \neq j} \kappa_{ji} x_i^a x_j^{a-1} + \eta_j x_j^{b-1} \right)^2 \\ &= \kappa_{jj}^2 x_j^{4a-2} + \sum_{i \neq j, \ell \neq j} \kappa_{ji} \kappa_{j\ell} x_i^a x_\ell^a x_j^{2a-2} + \eta_j^2 x_j^{2b-2} + 2 \sum_{i \neq j} \kappa_{jj} \kappa_{ji} x_i^a x_j^{3a-2} \\ &\qquad - 2 \sum_{i \neq j} \kappa_{ji} \eta_j x_i^a x_j^{a+b-2} - 2 \kappa_{jj} \eta_j x_j^{2a+b-2}. \end{aligned}$$

Thus, plugging this back in the definition of D, we can split D into a sum of six terms D_1 through D_6, each of which is a sum of terms of the form

$$\begin{aligned}
&\int_{\mathbb{R}_+^m} \prod_{k=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_k^{2a} + \eta_{0,k} \frac{x_k^b - 1}{b}\right) h_j(x_j) x_i^{\mathrm{pow}_i} x_\ell^{\mathrm{pow}_\ell} x_j^{\mathrm{pow}_j} \,\mathrm{d}\boldsymbol{x} \\
=&\prod_{k \neq j} \int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_k^{2a} + \eta_{0,k} \frac{x_k^b - 1}{b}\right) x_k^{\mathrm{pow}_k} \,\mathrm{d}x_k \\
&\qquad\qquad \times \int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b}\right) h_j(x_j) x_j^{\mathrm{pow}_j} \,\mathrm{d}x_j
\end{aligned}$$

times a constant involving $\mathbf{K}$ and η_j, where $\mathrm{pow}_k \geq 0$ for each $k \neq j$. We have thus decomposed the integral into a product of univariate integrals. Note that

$$\int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_i^{2a} + \eta_{0,i} \frac{x_i^b - 1}{b}\right) x_i^{\mathrm{pow}_i} \,\mathrm{d}x_i$$

is finite for all $\mathrm{pow}_i \geq 0$ regardless of whether b is nonzero, since we assumed $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ to lie in the parameter space with a finite normalizing constant. Indeed, if $b > 0$ then the terms in the exponential is a regular polynomial with positive degree and a negative leading term; if $b = 0$ then integrability follows from $\eta_{0,i} + \mathrm{pow}_i \geq \eta_{0,i} > -1$. Thus, we only need to consider the univariate integral that involve the x_j terms, namely

$$\int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b}\right) h_j(x_j) x_j^{\mathrm{pow}_j} \,\mathrm{d}x_j,$$

where pow_j takes value in $\{4a-2, 2a-2, 2b-2, 3a-2, a+b-2, 2a+b-2\} \subseteq [2\min\{a,b\} - 2, 4a-2]$. We split the integral into two parts over $[0,1]$ and $[1,\infty]$, respectively.

- If $b > 0$, on $[0,1]$ the exponential part is bounded above and below by positive constants, and for (A1) we require $h_j(x) = o(x^{1-\min\{a,b\}})$ as $x \searrow 0^+$, so the integrand is $o(x^{\min\{a,b\}-1}) = o(x^{-1})$ and is thus integrable on $[0,1]$. The integrand on $[1,\infty)$ is integrable as in (A1) we assume h to grow at most polynomially.
- If $b = 0$, $\mathrm{pow}_j \in [-2, 4a-2]$ and the integrand becomes

  $$\exp(-N_{\mathbf{K}_0} x_j^{2a}/(2a)) h_j(x_j) x_j^{\mathrm{pow}_j + \eta_{0,j}}.$$

  On $[0,1]$, (A1) requires $h_j(x) = o(x^{1-\min_j \eta_{0,j}})$, so $h_j(x_j) x_j^{\mathrm{pow}_j + \eta_{0,j}} = o(x^{-1})$ and the integrand is again integrable. Integrability on $[1,\infty)$ follows similarly to the case with $b > 0$.

Now consider the second part of (A2). By definition $\mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1$ equals

$$\int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^m \left| h_j'(X_j) \partial_j \log p(\boldsymbol{X}) + h_j(X_j) \partial_j^2 \log p(\boldsymbol{X}) \right| \mathrm{d}\boldsymbol{x}$$

$$\leq \sum_{j=1}^{m} \int_{\mathbb{R}_+^m} \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b}\right) \left| h_j'(x_j) \left(-\kappa_{jj} x_j^{2a-1} - \sum_{i \neq j} \kappa_{ji} x_i^a x_j^{a-1} + \eta_j x_j^{b-1}\right) \right.$$
$$\left. + h_j(x_j) \left(-\kappa_{jj}(2a-1) x_j^{2a-2} - \sum_{i \neq j} \kappa_{ji}(a-1) x_i^a x_j^{a-2} + (b-1)\eta_j x_j^{b-2}\right) \right| \mathrm{d}\boldsymbol{x}.$$

By the triangle inequality and the fact that $h_j \geq 0$ and $h_j' \geq 0$, similar to the proof for the first part, for each j the integral can be bounded by a sum of six integrals, each of the form

$$\text{const} \times \int_{\mathbb{R}_+^m} \prod_{k=1}^{m} \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_k^{2a} + \eta_{0,k} \frac{x_k^b - 1}{b}\right) h_j(x_j) x_i^{\mathrm{pow}_i} x_j^{\mathrm{pow}_j} \, \mathrm{d}\boldsymbol{x},$$

or with h_j replaced by h_j'. Finiteness thus follows from the same type of discussion by noting that $h_j(x) = o\left(x^{1-\min\{a,b\}}\right)$ and $h_j'(x) = o\left(x^{-\min\{a,b\}}\right)$.

We conclude that if the true and the proposed parameters give densities with finite normalizing constants, and if h satisfies assumption (A1), then (A2) is automatically satisfied.

In the centered case where we assume $\boldsymbol{\eta} \equiv \mathbf{0}$, we only need $\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{1-a} = 0$ as it is a special case with $b = 2a$. ■

A.4. Proof of Theorems in Section 6

Proof [Proof of Corollary 14] By Theorem 13, under assumptions in that theorem, the support of $\hat{\boldsymbol{\Psi}}$ is a subset of the true support of $\boldsymbol{\Psi}_0$, and $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda$. Since $\boldsymbol{\Psi}_0$ has $|S_0|$ nonzero entries,

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

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

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

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

Proof [Proof of Theorem 15] The proof is based on Theorem 13 and a probabilistic bound on $\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}} - \boldsymbol{\Gamma}_0\|_\infty$, where in the case of centered Gaussian $\boldsymbol{\Gamma} = \mathrm{diag}(\mathbf{x}\mathbf{x}^\top, \ldots, \mathbf{x}\mathbf{x}^\top)$. Denote $\boldsymbol{\Sigma}_0 = \mathbf{K}_0^{-1}$. In particular, given $\tau > 2$ we wish to show that for $\epsilon = 80\sqrt{2}c_0 \max_j(\Sigma_{0,jj})$, assuming $c_0 \equiv \sqrt{(\tau \log m + \log 4)/n} < 1/\sqrt{2}$,

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

and so the results follow from Theorem 13.

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

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

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 the complement of $\mathcal{E}_{j,j}(\epsilon/2)$, we have

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

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

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

Since $\tau > 2$, it holds that $1 + \left(1 + 1/(40\sqrt{2}c_0)\right)^{-1} = 2 - (1 + 40\sqrt{2}c_0)^{-1}$ is larger than $2 - (1 + 80\sqrt{\log m/n})^{-1} \equiv C(n, m)$, so it is safe to choose any $\delta \in (1, C(n, m))$. Thus by the requirement on ϵ, the theorem statement holds when $n > \max(c^* c_1^2 d_{\mathbf{K}}^2, 2)(\tau \log m + \log 4)$ with $c^* = 12800\max_j(\Sigma_{0,jj})^2$. ■

Proof [Proof of Theorem 16] The proof of Theorem 13 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 13 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 (21) that the j^{th} block of $\mathbf{\Gamma}_{\boldsymbol{\gamma}} \in \mathbb{R}^{m^2 \times m^2}$ has (k, ℓ)-th entry

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) + \gamma_{kj}\mathbb{1}_{\{k=\ell\}} - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| > \epsilon_1\right) \\
\leq &\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| > \epsilon_1/2\right) \qquad (39) \\
\leq &2\exp\left(-\min\left(\frac{n\epsilon_1^2}{8M^2 c_{\boldsymbol{X}}^4}, \frac{n\epsilon_1}{4Mc_{\boldsymbol{X}}^2}\right)\right) \leq \frac{1}{3m^\tau}. \qquad (40)
\end{aligned}$$

Denote the event inside the probability in (39) as $\mathcal{E}_{k,\ell,j}(\epsilon_1/2)$.
By definition,

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

By Lemmas 21.2 and 22.1 from Appendix B, $\mathrm{var}(X_k) \leq \Sigma_{kk}$, 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, setting $\epsilon_1 = 2Mc_{\boldsymbol{X}}^2 c_{n,m}$, on the complement of $\mathcal{E}_{k,k,j}(\epsilon_1/2)$ we have

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

Then

$$\frac{1}{1 + 1/(4ec_{n,m})}\frac{1}{n}\sum_{i=1}^n {X_k^{(i)}}^2 h_j\left(X_j^{(i)}\right) \leq \epsilon_1/2 \qquad (41)$$

on the complement of $\mathcal{E}_{k,k,j}(\epsilon_1/2)$, again with $c_{n,m} \equiv \max\left\{\frac{2(\log m^\tau + \log 6)}{n}, \sqrt{\frac{2(\log m^\tau + \log 6)}{n}}\right\}$.
Note that the multiplier on the left of (41) is increasing in $c_{n,m}$, and that $2(\log m^\tau + \log 6) > 6\log m$ by the assumption that $\tau > 3$. Thus, if we let

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

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

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

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

$$\mathbb{P}\left(\max_{j,k,\ell}\left|n^{-1}\sum_{i=1}^n X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) + \gamma_{kj}\mathbb{1}_{\{k=\ell\}} - \mathbb{E}_0 X_k X_\ell h_j(X_j)\right| \geq \epsilon_1\right) \leq \frac{1}{3m^{\tau-3}}. \quad (42)$$

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

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

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

$$\mathbb{P}\left(\max_{j,k}\left|n^{-1}\sum_{i=1}^{n} X_k^{(i)} h_j'\left(X_j^{(i)}\right) - \mathbb{E}_0 X_k h_j'(X_j)\right| \geq \epsilon_{2,1}\right) \leq \frac{1}{3m^{\tau-3}}, \tag{43}$$
$$\mathbb{P}\left(\max_{j}\left|n^{-1}\sum_{i=1}^{n} h_j\left(X_j^{(i)}\right) - \mathbb{E}_0 h_j(X_j)\right| \geq \epsilon_{2,2}\right) \leq \frac{1}{3m^{\tau-3}}. \tag{44}$$

Hence, by (42) (43) (44), 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

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

Then $d_{\mathbf{K}_0}\epsilon_1 \leq \alpha/(6c_{\boldsymbol{\Gamma}_0})$ and the results follow from Theorem 13. ■

Proof [Proof of Theorem 17] Similar to the proof of Theorem 16, by Theorem 13 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

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

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

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

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

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

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

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

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

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

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 16, and use union bounds to arrive at analogous consistency results, modulus different constants. The amplifiers $\boldsymbol{\gamma}$ can be incorporated analogously. ■

Appendix B. Auxiliary 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.

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

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

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

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

The definition of sub-Gaussian norm here allows for a non-centered variable and differs from the one in Vershynin (2012), which uses $\|X\|_q$. Instead, it coincides with θ_2 in Buldygin and Kozachenko (2000). The sub-Gaussian parameter is defined as in Buldygin and Kozachenko (2000) and the sub-exponential norm as in Vershynin (2012).

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; so* $\tau(X+Y) \leq \tau(X) + \tau(Y)$*. 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,* $\tau\left(n^{-1}\sum_{i=1}^n X_i\right) \leq n^{-1/2}\tau(X_i)$*.*

3) If X_1 *and* X_2 *are sub-Gaussian (not necessarily independent) 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 and* $q > 0$*, then*

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

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

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

6) (Vershynin, 2012) 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* $\epsilon > 0$*,*

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

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 $\epsilon > 0$*,*

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

Proof
1) For $r = 1, 2$, by the 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),

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

4-6) These are Lemma 1.4 and Theorem 1.5 in Buldygin and Kozachenko (2000), and a consequence of Corollary 5.17 in Vershynin (2012).

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

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

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 $\epsilon > 0$,*

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

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

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_j^{(i)}h_0\left(X_k^{(i)}\right) - \mathbb{E}X_jh_0(X_k)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8M_0^2(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j)^2}\right),$$

$$\tau\left(n^{-1}\sum_{i=1}^n X_j^{(i)}h_0\left(X_k^{(i)}\right) - \mathbb{E}X_jh_0(X_k)\right) \leq \frac{2M_0}{\sqrt{n}}\left(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j\right),$$

$$\left\|n^{-1}\sum_{i=1}^n X_j^{(i)}h_0\left(X_k^{(i)}\right) - \mathbb{E}X_jh_0(X_k)\right\|_{\psi_2} \leq \frac{4M_0}{\sqrt{en}}\left(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j\right).$$

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

$$\|X_jX_kh_0(X_\ell) - \mathbb{E}X_jX_kh_0(X_\ell)\|_{\psi_1} \leq \frac{M_0}{2e}c_{\boldsymbol{X}}^2, \tag{45}$$

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

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

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 $t \in \mathbb{R}$,

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

which is equivalent to

$$t^2\Sigma_{11}/2 + t\mathbb{E}X_1 - \log\mathbb{E}\exp(tX_1) \geq 0 \quad \forall t \in \mathbb{R}. \tag{46}$$

Since the left-hand side of (46) equals 0 at $t = 0$, it suffices to show that its derivative,

$$t\Sigma_{11} + \mathbb{E}X_1 - \frac{\mathrm{d}\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t} = t\Sigma_{11} + \mathbb{E}X_1 - \frac{\frac{\mathrm{d}\mathbb{E}\exp(tX_1)}{\mathrm{d}t}}{\mathbb{E}\exp(tX_1)}, \tag{47}$$

is non-negative on $(0, \infty)$ and non-positive on $(-\infty, 0)$. By properties of moment-generating functions, $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\exp(tX_1)$ evaluated at $t = 0$ equals $\mathbb{E}X_1$, so (47) equals 0 at $t = 0$. It in turn suffices to show the derivative of (47), namely

$$\Sigma_{11} - \frac{\mathrm{d}^2\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t^2} \tag{48}$$

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

Given any vector $\boldsymbol{v} \in \mathbb{R}^p$, define $\mathbb{R}^p_+ - \boldsymbol{v} \equiv \{\boldsymbol{u} - \boldsymbol{v} : \boldsymbol{u} \in \mathbb{R}^p_+\}$. By Tallis (1961), denoting the first column of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_1$, the moment-generating function of the marginal distribution of X_1 is

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

(48) thus becomes

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

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

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

with $\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)$ log-concave in $\boldsymbol{x}$ and $\mathbb{1}_{\mathbb{R}^p_+ - \boldsymbol{\mu}}(\boldsymbol{x} + t\boldsymbol{\Sigma}_1)$ log-concave in $(\boldsymbol{x}, t)$ since $\mathbb{R}^p_+ - \boldsymbol{\mu}$ is a convex set. 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) of the Lemma 21,

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

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

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

Since h_0 is a function bounded by M_0, by definition

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

Then by 1) of Lemma 21 again,

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

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

Although not used for our consistency results, in the special case of $h_0 \equiv 1$, we also have the following lemma. The notable difference between bounds (49) below and (45) 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 (45) is smaller and thus gives a tighter bound, but (49) 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 $j, k \in \{1, \ldots, p\}$, for any $\epsilon > 0$,*

$$\mathbb{P}\left(n^{-1}\left|\sum_{i=1}^{n} X_j^{(i)} X_k^{(i)} - \mathbb{E}X_j X_k\right| \geq \epsilon\right) \leq 4\exp\left(-\min\left(\frac{2n\epsilon^2}{C_1^2}, \frac{n\epsilon}{C_1}\right)\right), \tag{49}$$

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 $\mathbb{E}X_j$ may be nonzero 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

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

We next define

$$\begin{aligned}
&Z_{jk}^{(i)} \equiv {U_{jk}^{(i)}}^2 - \mathbb{E}U_{jk}^2 = {A_{jk}^{(i)}}^2 + B_{jk}^{(i)} + C_{jk}, && \overline{X}_j^{(i)} \equiv X_j^{(i)} - \mathbb{E}X_j, \\
&A_{jk}^{(i)} \equiv \overline{X}_j^{(i)} + \overline{X}_k^{(i)}, \quad B_{jk}^{(i)} \equiv 2(\mathbb{E}X_j + \mathbb{E}X_k)(\overline{X}_j^{(i)} + \overline{X}_k^{(i)}), && C_{jk} \equiv -\mathbb{E}(\overline{X}_j^{(i)} + \overline{X}_k^{(i)})^2.
\end{aligned}$$

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 $\leq 2(\mathbb{E}X_j + \mathbb{E}X_k)\left(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}\right)$. 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 $q \geq 2$

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

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 1) of Lemma 22, we then have

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

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

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

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

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

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

Appendix C. Simulation Results for Erdös-Rényi Graphs

We revisit the simulations from Section 7 but use Erdös-Rényi (ER) graphs in which each possible edge is independently included with probability π. Independent uniform draws from $[0.5, 1]$ are used to fill the non-zero off-diagonal entries of the symmetric matrix $\mathbf{K}_0$. The diagonal elements are set such that $\mathbf{K}_0$ has minimum eigenvalue 0.1. We choose $\pi = 0.08$ for $n = 1000$, and $\pi = 0.02$ for $n = 80$.

C.1. Truncated GGMs

In this section we present the results for truncated GGMs.

C.1.1. Choice of h

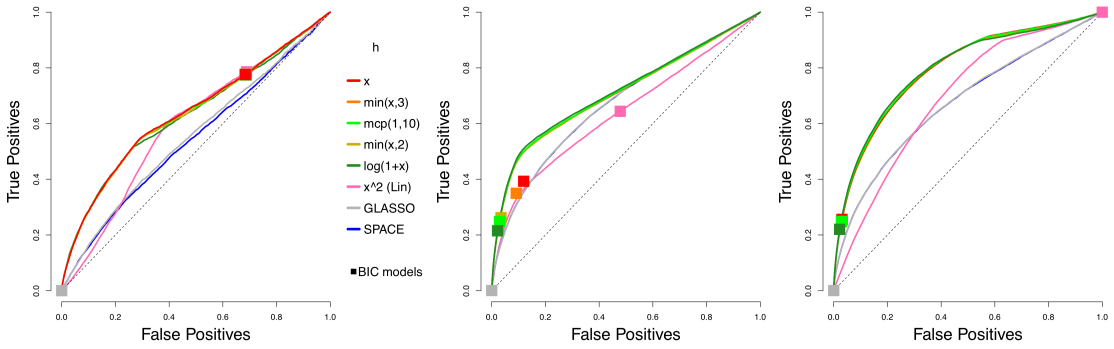
The results for truncated centered GGMs are reported in Table 4 and Figure 13. Those for truncated non-centered GGMs using the profiled estimator are in Table 5 and Figure 14.

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

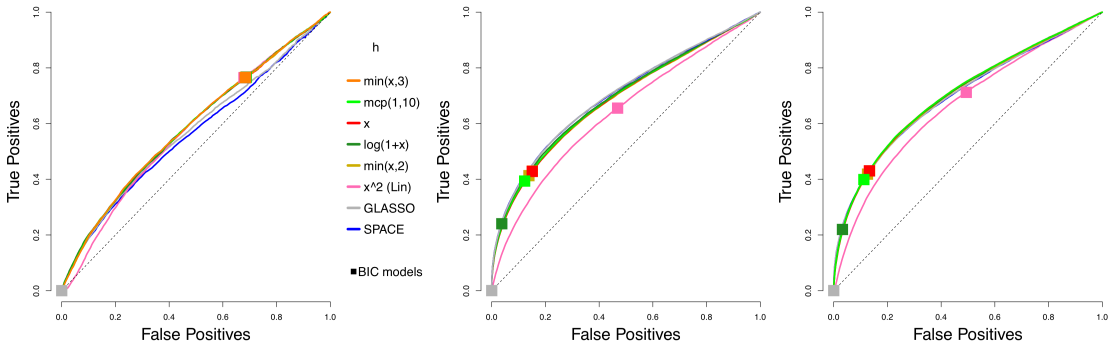
Figure 13: Average ROC curves of our *centered* estimator for $m = 100$ variables and two sample sizes n under various choices of h, compared to SPACE and GLASSO, for the *truncated centered GGM* case. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.

Non-centered profiled, $n = 80$, multiplier 1.8647, 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 as for Table 4.

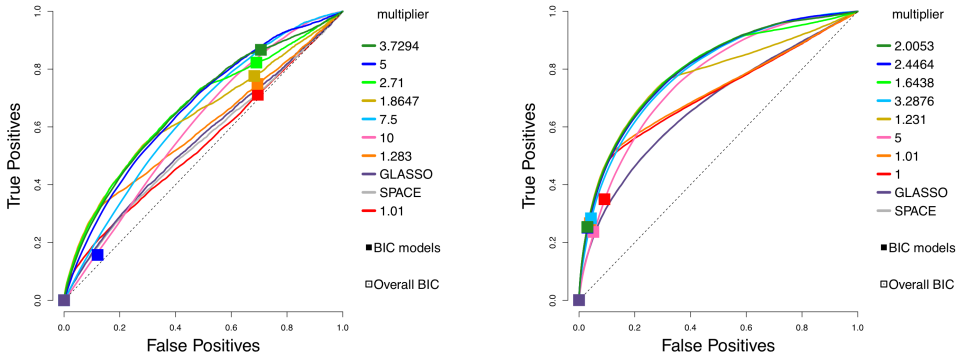


(a) $n = 80$, mult $= 1.8647$, ER (b) $n = 1000$, mult $= 1$, ER (c) $n = 1000$, mult $= 1.6438$, ER

Figure 14: Average for the truncated non-centered GGM case. $n = 80$ or 1000, $m = 100$.

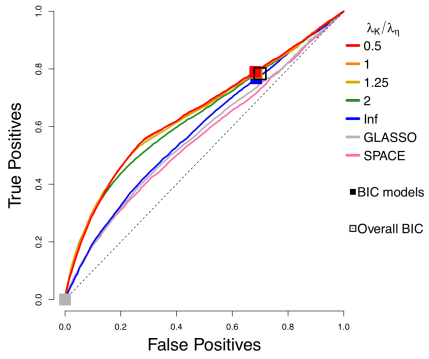
C.1.2. Choice of multiplier

The results for truncated centered GGMs where each curve represents a different multiplier are shown in Figure 15, and those for truncated non-centered GGMs are in Figure 16, where each curve corresponds to a different ratio $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$.

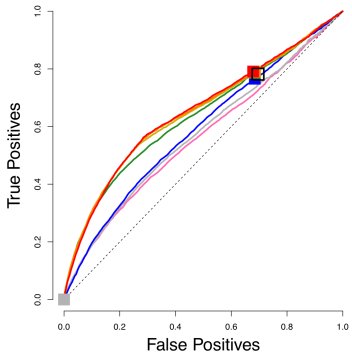


(a) $n = 80$, ER (b) $n = 1000$, ER

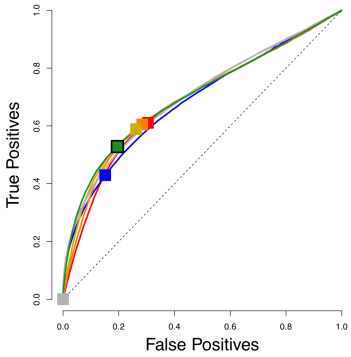
Figure 15: Performance of $\min(x, 3)$ for truncated centered GGMs with different multipliers, compared to GLASSO and SPACE, in the centered setting, $n = 80$ or 1000.



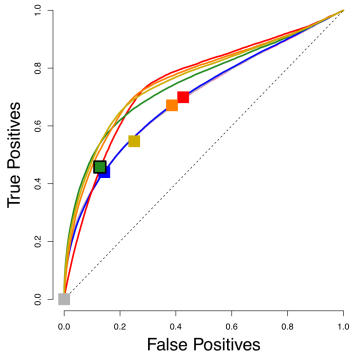
(a) $n = 80$, mult $= 1.7897$, ER



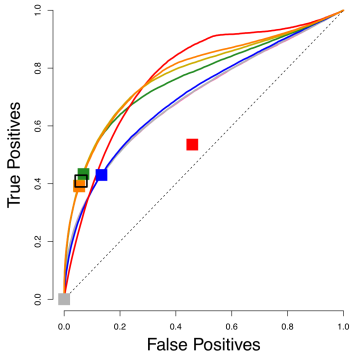
(b) $n = 80$, mult $= 1.8647$, ER



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



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

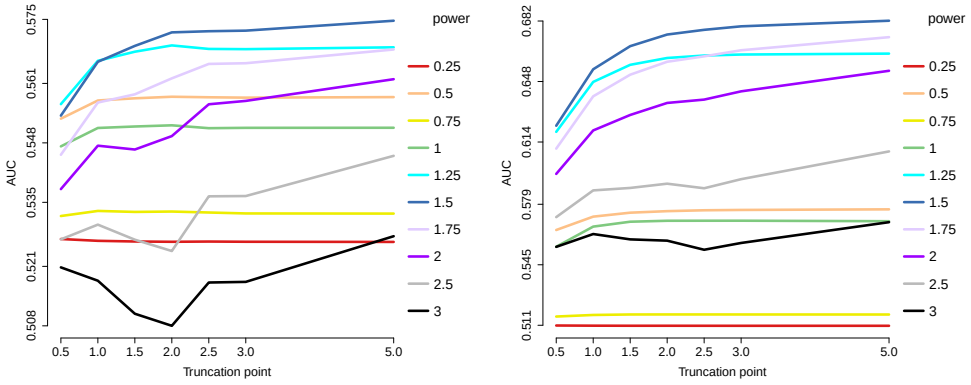


(e) $n = 1000$, mult $= 1.6438$, ER

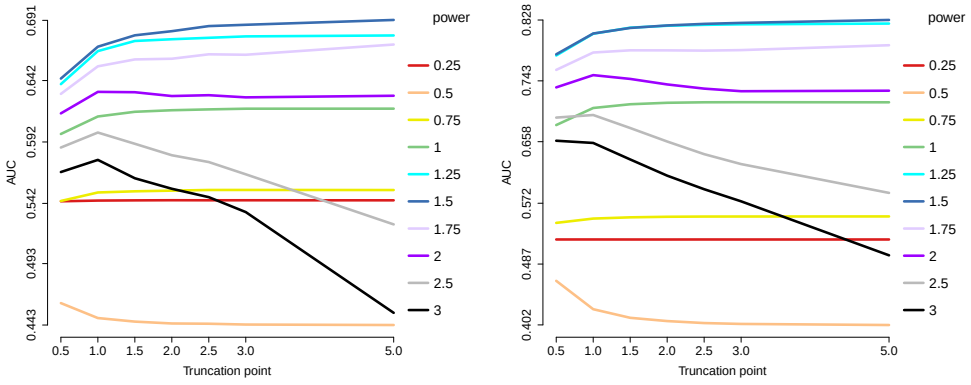
Figure 16: Performance of the non-centered estimator with $h(x) = \min(x, 3)$. Each curve corresponds to a different choice of $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$. Squares indicate models picked by eBIC with refit. The square with black outline has the highest eBIC among all models (combinations 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.

C.2. Other a/b Models

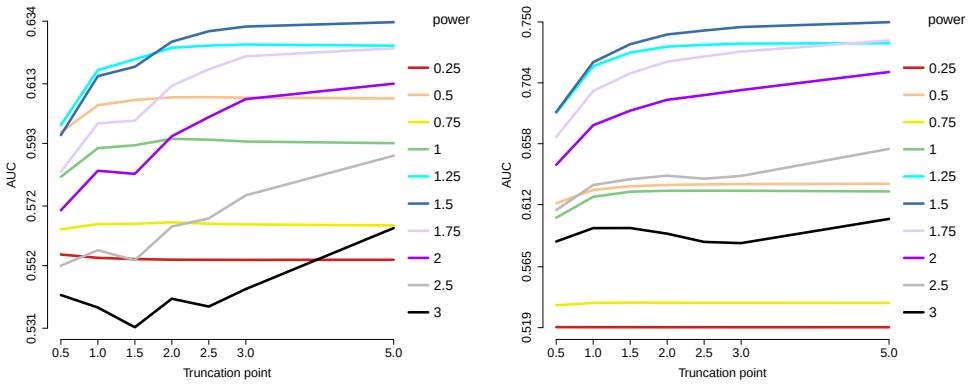
Figure 17 exhibits the results for the exponential models.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est, ER (b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est, ER



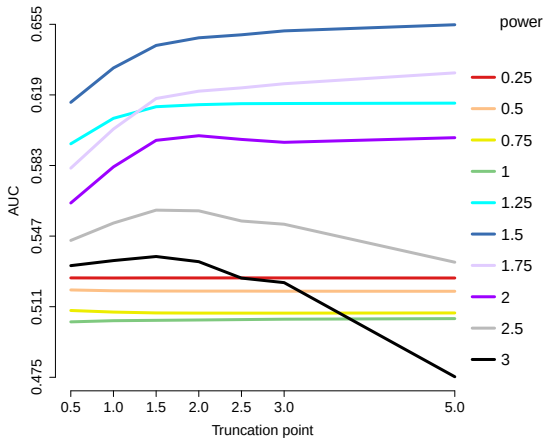
(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered est, ER (d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered est, ER



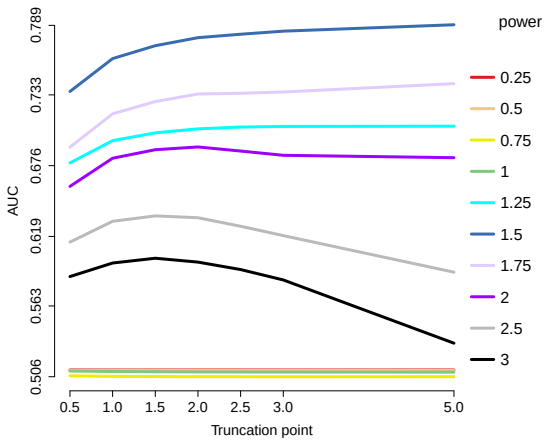
(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est, ER (f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est, ER

Figure 17: AUCs for edge recovery using generalized score matching for the exponential models. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.

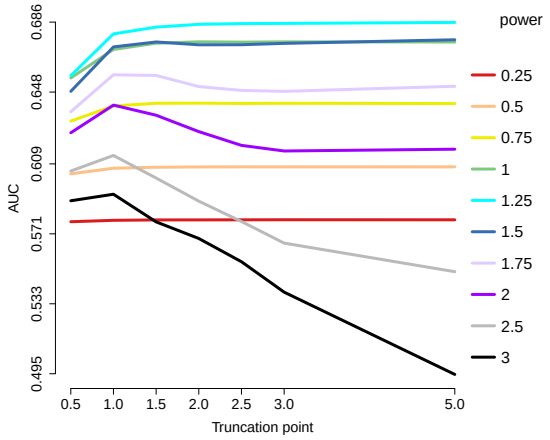
Figure 18 displays the results for the gamma models.



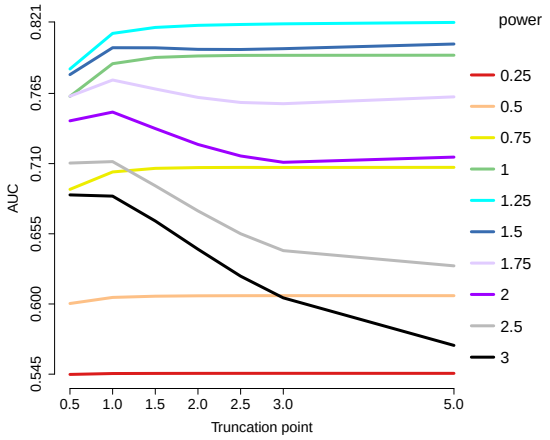
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator, ER



(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator, ER



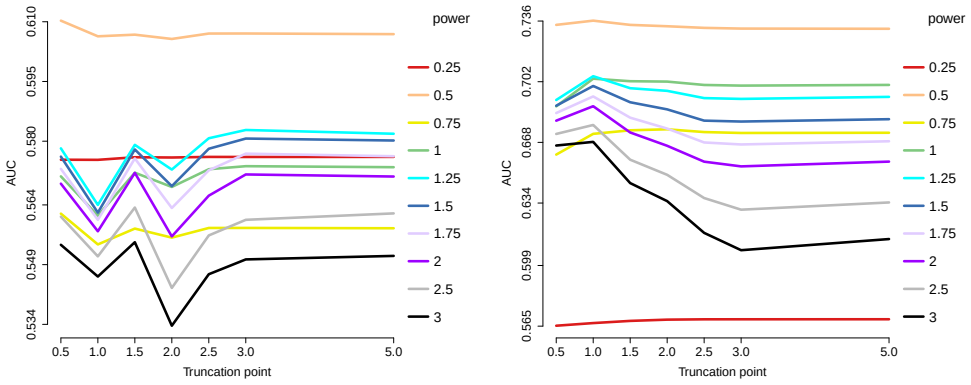
(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator, ER



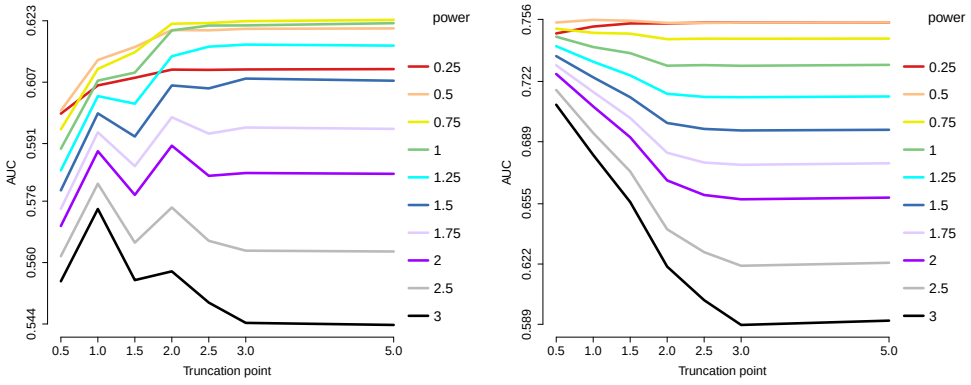
(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator, ER

Figure 18: AUCs for edge recovery using generalized score matching for the gamma models. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.

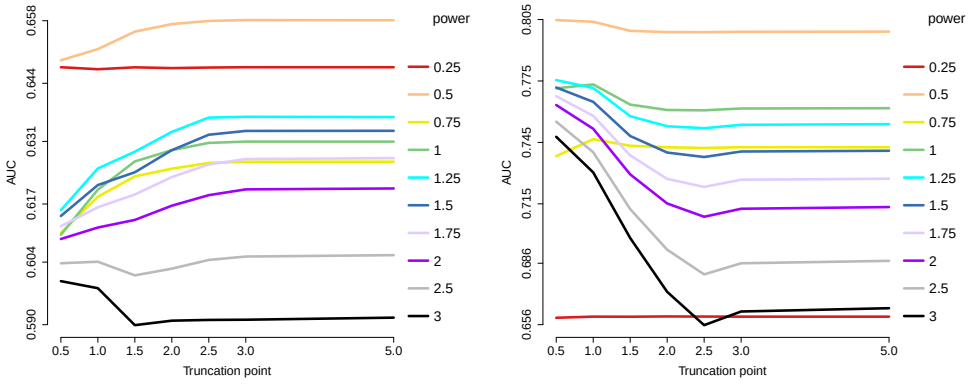
Figures 19 and 20 demonstrate the results for $a = 3/2$ and $b = 1/2$ or $b = 0$, respectively.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est, ER (b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est, ER

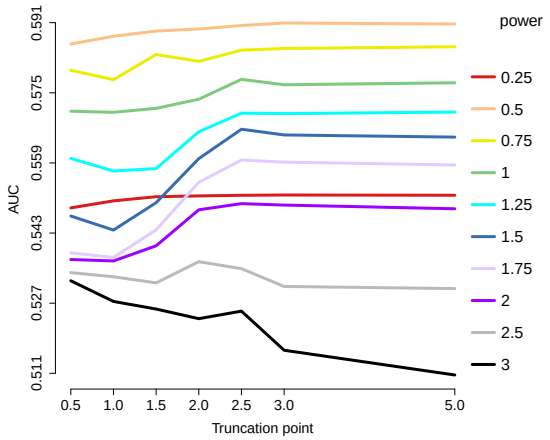


(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered est, ER (d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered est, ER

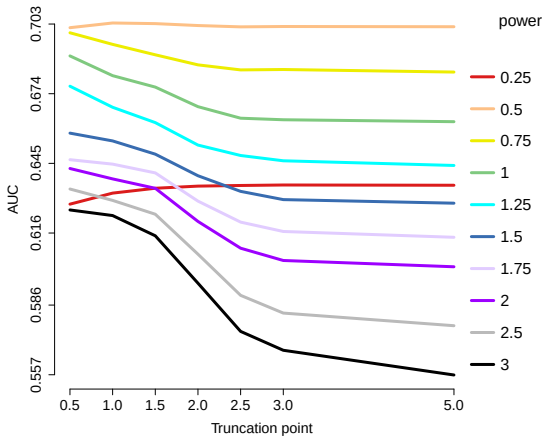


(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est, ER (f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est, ER

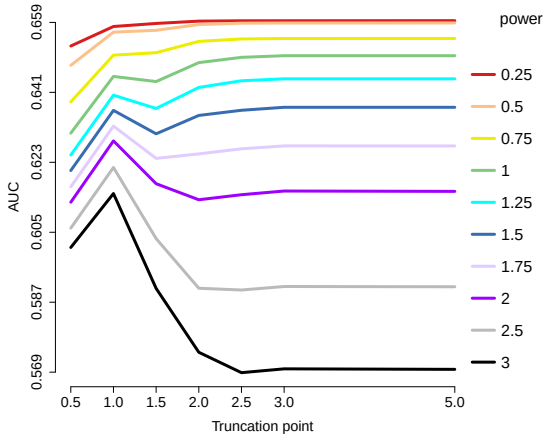
Figure 19: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 1/2$. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.



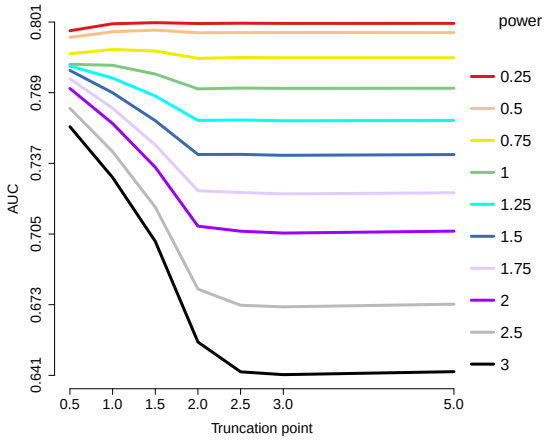
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator, ER



(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator, ER



(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator, ER



(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator, ER

Figure 20: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 0$. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.

References

Murilo P. Almeida and Basilis Gidas. A variational method for estimating the parameters of MRF from complete or incomplete data. *Ann. Appl. Probab.*, 3(1):103–136, 1993.

Rina Foygel Barber and Mathias Drton. High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.*, 9(1):567–607, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities*. Oxford University Press, Oxford, 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.

V. V. Buldygin and Yu. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000. Translated from the 1998 Russian original by V. Zaiats.

Scott L. Carter, Christian M. Brechbühler, Michael Griffin, and Andrew T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

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, 2015.

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.*, 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. *J. Amer. Statist. Assoc.*, 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. *Comput. Statist. Data Anal.*, 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, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

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

Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha JM. Walhout, Michael E. Cusick, Frederick P. Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88, 2004.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Comput. Statist. Data Anal.*, 51(5): 2499–2512, 2007.

David Inouye, Pradeep Ravikumar, and Inderjit Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2445–2453, 2016.

Hawoong Jeong, Sean P. Mason, A-L. Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(4):803–825, 2015.

Steffen L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996.

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

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

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *J. Multivariate Anal.*, 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. *Ann. Statist.*, 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. Extensive scoring rules. *Electron. J. Stat.*, 10(1):1098–1108, 2016.

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

Jie Peng, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 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. *Ann. Statist.*, 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. *Electron. J. Stat.*, 5:935–980, 2011.

G. M. Tallis. The moment generating function of the truncated multi-normal distribution. *J. Roy. Statist. Soc. Ser. B*, 23:223–229, 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. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 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.

Hannu Väliaho. Criteria for copositive matrices. *Linear Algebra Appl.*, 81:19–34, 1986.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

Eunho Yang, Pradeep Ravikumar, Genevera I. Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, 16:3813–3847, 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.

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

Małgorzata Żak-Szatkowska and Małgorzata Bogdan. Modified versions of the Bayesian information criterion for sparse generalized linear models. *Comput. Statist. Data Anal.*, 55(11):2908–2924, 2011.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 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. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.