Graphical Models for Non-Negative Data Using Generalized Score Matching

Shiqing Yu
University of Washington

Mathias Drton
University of Washington

Ali Shojaie
University of Washington

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. In contrast, 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. We also generalize the regularized score matching method of Lin et al. (2016) for non-negative Gaussian graphical models, with improved theoretical guarantees.

1 INTRODUCTION

Graphical models (Lauritzen, 1996) characterize the relationships among random variables $(X_i)_{i \in V}$ indexed by the nodes of a graph $G = (V, E)$; here, $E \subseteq V \times V$ is the set of edges in G. When the graph G is undirected, two variables X_i and X_j are required to be conditionally independent given all other $(X_k)_{k \in V \setminus \{i,j\}}$ if there is no edge between i and j. The smallest graph G such that this property holds is called the *conditional independence graph* of the random vector $\boldsymbol{X} \equiv (X_i)_{i \in V}$. See Drton and Maathuis (2017) for a more detailed introduction to these and other graphical models.

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 known as the *concentration matrix*. More specifically, X_i and X_j are conditionally independent given all other variables in $\boldsymbol{X}$ if and only if the (i, j)-th and the (j, i)-th entry of $\mathbf{K}$ are both zero. This simple relation underlies a rich literature on GGMs, including Drton and Perlman (2004), Meinshausen and Bhlmann (2006), Yuan and Lin (2007) and Friedman et al. (2008), among others.

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 and other exponential family models (Ravikumar et al., 2010; Chen et al., 2014; Yang et al., 2015), as well as semi- or non-parametric estimation techniques (Fellinghauer et al., 2013; Voorman et al., 2013). In this paper, we focus on non-negative Gaussian random variables, as recently considered by Lin et al. (2016) and Yu et al. (2016).

The *probability density function* of a *non-negative Gaussian* random vector $\boldsymbol{X}$ is proportional to that of the corresponding Gaussian vector, but restricted to the non-negative orthant. More specifically, let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be a mean vector and a covariance matrix for an m-variate random vector, respectively. Then $\boldsymbol{X}$ follows a truncated normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ if it has density $\exp\left(-0.5(\boldsymbol{x} - \boldsymbol{\mu})^\top \mathbf{K}(\boldsymbol{x} - \boldsymbol{\mu})\right)$ on $\mathbb{R}^m_+ \equiv [0, +\infty)^m$, where $\mathbf{K} \equiv \boldsymbol{\Sigma}^{-1}$ is the inverse covariance parameter. We denote this as $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{K})$. The conditional independence graph of a truncated normal vector is determined by $\mathbf{K} \equiv [\kappa_{ij}]_{i,j}$ just as in the Gaussian case: X_i and X_j are conditionally independent given all other variables if $\kappa_{ij} = \kappa_{ji} = 0$.

Suppose $\boldsymbol{X}$ is a continuous random vector with distribution P_0, density p_0 wrt Lebesgue measure, and support $\mathbb{R}^m$, so $p_0(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \mathbb{R}^m$.

Let $\mathcal{P}$ be a family of distributions with twice differentiable densities that we know only up to a (possibly intractable) normalizing constant. The SM estimator of p_0 using $\mathcal{P}$ as a model is the min-

imizer of the expected squared ℓ_2 distance between the gradients of $\log p_0$ and a log-density from $\mathcal{P}$. Formally, we minimize the loss from (1) below. Although the loss depends on p_0, partial integration can be used to rewrite it in a form that can be approximated by averaging over the sample without knowing p_0. The key advantage of score matching is that the normalizing constant cancels from the gradient of log-densities. Furthermore, for exponential families, the loss is quadratic in the parameter of interest, making optimization straightforward.

When dealing with distributions supported on a proper subset of $\mathbb{R}^m$, the partial integration arguments underlying the score matching estimator may fail due to discontinuities at the boundary of the support. To circumvent this problem, Hyvärinen (2007) introduced a *modified score matching estimator* for data supported on $\mathbb{R}_+^m$ by minimizing a loss in which boundary effects are dampened by multiplying gradients elementwise with the identity functions x_j; see (3) below. Lin et al. (2016) estimate truncated GGMs based on this modification, with an ℓ_1 penalty on the entries of $\mathbf{K}$ added to the loss. In this paper, we show that elementwise multiplication with functions other than x_j can lead to improved estimation accuracy in both simulations and theory. Following Lin et al. (2016), we will then use the proposed GSM framework to estimate the matrix $\mathbf{K}$.

Section 2 introduces score matching and our proposed *generalized score matching*. In Section 3, we apply generalized score matching to exponential families, with univariate truncated Gaussian distributions as an example. *Regularized generalized score matching* for graphical models is formulated in Section 4. Simulation results are given in Section 5.

1.1 Notation

Subscripts are used to refer to entries in vectors and columns in matrices. Superscripts are used to refer to rows in matrices. For example, when considering a matrix of observations $\mathbf{x} \in \mathbb{R}^{n \times m}$, each row being a sample of m measurements/features for one observation/individual, $X_j^{(i)}$ is the j-th feature for the i-th observation. For a rv $\boldsymbol{X}$, X_j refers to its j-th component.

The vectorization of a matrix $\mathbf{K} = [\kappa_{ij}]_{i,j} \in \mathbb{R}^{q \times r}$ is obtained by stacking its columns:

$$\text{vec}(\mathbf{K}) = (\kappa_{11}, \ldots, \kappa_{q1}, \kappa_{12}, \ldots, \kappa_{q2},$$

$$\ldots, \kappa_{1r}, \ldots, \kappa_{qr})^\top.$$

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

noted $\|\boldsymbol{v}\|_a = (\sum_{j=1}^m |v_j|^q)^{1/a}$, and the ℓ_∞-norm is defined as $\|\boldsymbol{v}\|_\infty = \max_{j=1,\ldots,q} |v_j|$. The ℓ_a-ℓ_b operator norm for matrix $\mathbf{K} \in \mathbb{R}^{q \times r}$ is written as $\|\|\mathbf{K}\|\|_{a,b} \equiv \max_{\boldsymbol{x} \neq \mathbf{0}} \|\mathbf{K}\boldsymbol{x}\|_b / \|\boldsymbol{x}\|_a$ with shorthand notation $\|\|\mathbf{K}\|\|_a \equiv \|\|\mathbf{K}\|\|_{a,a}$. By definition, $\|\|\mathbf{K}\|\|_\infty \equiv \max_{i=1,\ldots,q} \sum_{j=1}^r |\kappa_{ij}|$. We write the Frobenius norm of $\mathbf{K}$ as $\|\|\mathbf{K}\|\|_F \equiv \|\text{vec}(\mathbf{K})\|_2 \equiv (\sum_{i=1}^q \sum_{j=1}^r \kappa_{ij}^2)^{1/2}$, and its max norm $\|\mathbf{K}\|_\infty \equiv \|\text{vec}(\mathbf{K})\|_\infty \equiv \max_{i,j} |\kappa_{ij}|$.

For a scalar function f, we define $\partial_j f(\boldsymbol{x})$ as its partial derivative with respect to the j-th component evaluated at x_j, and $\partial_{jj} f(\boldsymbol{x})$ the corresponding second partial derivative. For vector-valued $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$, $\boldsymbol{f}(x) = (f_1(x), \ldots, f_m(x))^\top$, let $\boldsymbol{f}'(x) = (f_1'(x), \ldots, f_m'(x))^\top$ be the vector of derivatives and $\boldsymbol{f}''(x)$ likewise.

Throughout the paper, $\mathbf{1}_n$ refers to a vector of all 1's of length n. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{a} \circ \boldsymbol{b} \equiv (a_1 b_1, \ldots, a_m b_m)^\top$. Moreover, when we speak of the "density" of a distribution, we mean its probability density function w.r.t. Lebesgue measure. When it is clear from the context, $\mathbb{E}_0$ denotes the expectation under the true distribution.

2 SCORE MATCHING

2.1 Original Score Matching

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

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x}. \quad (1)$$

The gradients in (1) can be thought of as gradients with respect to a hypothetical location parameter, evaluated at $\mathbf{0}$ (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, (1) can be rewritten as

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x})$$

$$\sum_{j=1}^m \left[\partial_{jj} \log p(\boldsymbol{x}) + \frac{(\partial_j \log p(\boldsymbol{x}))^2}{2} \right] \mathrm{d}\boldsymbol{x} \quad (2)$$

plus a constant independent of p. Clearly, the integral in (2) can be approximated by its corresponding sample average without knowing the true density p_0, and can thus be used to estimate p_0.

2.2 Score Matching for Non-Negative Data

When the true density p_0 is only supported on a proper subset of $\mathbb{R}^m$, the integration by parts underlying the equivalence of (1) and (2) 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*

$$J_+(P) = \int_{\mathbb{R}^m_+} p_0(\boldsymbol{x}) \| \nabla \log p(\boldsymbol{x}) \circ \boldsymbol{x} \\ - \nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{x} \|_2^2 \, \mathrm{d}\boldsymbol{x}. \quad (3)$$

This loss can be motivated by considering gradients of the true and model log-densities w.r.t. a hypothetical scale parameter (Hyvärinen, 2007). Under regularity conditions, it can again be rewritten as the expectation (under P_0) of a function independent of p_0, thus allowing one to estimate p_0 by minimizing the corresponding sample loss.

2.3 Generalized Score Matching for Non-Negative Data

We consider the following generalization of the non-negative score matching loss (3).

Definition 1. Suppose random vector $\boldsymbol{X} \in \mathbb{R}^m_+$ has true distribution P_0 with density p_0 that is twice differentiable and supported on $\mathbb{R}^m_+$. Let $\mathcal{P}_+$ be the family of all distributions with twice differentiable densities supported on $\mathbb{R}^m_+$, and suppose $P_0 \in \mathcal{P}_+$. Let $h_1, \ldots, h_m : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a.e. positive functions that are differentiable almost everywhere, 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

$$J_{\boldsymbol{h}}(p) = \int_{\mathbb{R}^m_+} \frac{1}{2} p_0(\boldsymbol{x}) \| \nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} \\ - \nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}(\boldsymbol{x})^{1/2} \|_2^2 \, \mathrm{d}\boldsymbol{x}, \quad (4)$$

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

Choosing all $h_j(x) = x^2$ recovers the loss from (3). The key intuition for our generalized score matching is that we keep the h_j increasing but instead focus on functions that are bounded or grow rather slowly. This will result in reliable higher moments, leading to better practical performance and improved theoretical guarantees. We note that our approach could also be presented in terms of transformations of data; compare to Section 11 in Parry et al. (2012). In particular, log-transforming positive data into all of $\mathbb{R}^m$ and then applying (1) is equivalent to (3).

We will consider the following assumptions:

(A1) $p_0(\boldsymbol{x}) h_j(x_j) \partial_j \log p(\boldsymbol{x}) \rightarrow 0$ as $x_j \nearrow +\infty$

and as $x_j \searrow 0^+, \forall \boldsymbol{x}_{-j} \in \mathbb{R}^{m-1}_+, \forall p \in \mathcal{P}_+$,

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

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

where $\forall p \in \mathcal{P}_+$ is a shorthand for "for all p being the density of some $P \in \mathcal{P}_+$", and the prime symbol denotes component-wise differentiation.

Assumption (A1) validates integration by parts and (A2) ensures the loss to be finite. We note that (A1) and (A2) are easily satisfied when we consider exponential families with $\lim_{x \searrow 0^+} h_j(x) = 0$.

The following theorem states that we can rewrite $J_{\boldsymbol{h}}$ as an expectation (under P_0) of a function that does not depend on p_0, similar to (2).

Theorem 2. *Under (A1) and (A2),*

$$J_{\boldsymbol{h}}(p) = C + \int_{\mathbb{R}^m_+} p_0 \sum_{j=1}^m \big[h_j'(x_j) \partial_j (\log p(\boldsymbol{x})) \quad (5)$$

$$+ h_j(x_j) \partial_{jj} (\log p(\boldsymbol{x})) + \frac{1}{2} h_j(x_j) (\partial_j (\log p(\boldsymbol{x})))^2 \big] \, \mathrm{d}\boldsymbol{x},$$

where C is 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 (5) as

$$\hat{J}_{\boldsymbol{h}}(p) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \bigg\{ h_j'(X_j^{(i)}) \partial_j (\log p(\boldsymbol{X}^{(i)})) +$$

$$h_j(X_j^{(i)}) \bigg[\partial_{jj} (\log p(\boldsymbol{X}^{(i)})) + \frac{1}{2} \left(\partial_j (\log p(\boldsymbol{X}^{(i)})) \right)^2 \bigg] \bigg\}.$$

We first clarify estimation consistency, in analogy to Corollary 3 in Hyvärinen (2005).

Theorem 3. *Consider a model $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \subset \mathcal{P}_+$ with parameter space $\boldsymbol{\Theta}$, and suppose that the true data-generating distribution $P_0 \equiv P_{\boldsymbol{\theta}_0} \in \mathcal{P}_+$ with density $p_0 \equiv p_{\boldsymbol{\theta}_0}$. Assume that $P_{\boldsymbol{\theta}} = P_0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then the generalized $\boldsymbol{h}$-score matching estimator $\hat{\boldsymbol{\theta}}$ obtained by minimization of $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ over $\boldsymbol{\Theta}$ converges in probability to $\boldsymbol{\theta}_0$ as the sample size n goes to infinity.*

3 EXPONENTIAL FAMILIES

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

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

It is not difficult to show that under assumptions (A1) and (A2) from Section 2.3, the empirical generalized $\boldsymbol{h}$-score matching loss $\hat{J}_{\boldsymbol{h}}$ above can be rewritten as

$$\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^{\top}\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^{\top}\boldsymbol{\theta} + \text{const}, \qquad (6)$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{r^2}$ and $\boldsymbol{g} \in \mathbb{R}^r$ are sample averages of functions of the data matrix $\mathbf{x}$ only; the detailed expressions are omitted here.

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 4. *Suppose that*

(C1) $\boldsymbol{\Gamma}$ *is invertible a.s., and*

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

Then the minimizer of (6) is a.s. unique with solution $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$, and

$$\hat{\boldsymbol{\theta}} \rightarrow_{a.s.} \boldsymbol{\theta}_0 \quad and \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightarrow_d \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Gamma}_0^{-1}\right).$$

We note that (C1) holds if $h_j(X_j) > 0$ a.e. and $[\partial_j\boldsymbol{t}(\boldsymbol{X}^{(1)}), \ldots, \partial_j\boldsymbol{t}(\boldsymbol{X}^{(n)})] \in \mathbb{R}^{r \times n}$ has rank r a.e. for some $j = 1, \ldots, m$.

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

Example 5. Consider univariate ($m = r = 1$) truncated Gaussian distributions with unknown mean parameter μ, known variance parameter σ^2, so

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

Then, given i.i.d. samples $X_1, \ldots, X_n \sim p_{\mu_0}$, the generalized h-SM estimator of μ is

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

If $\lim_{x \searrow 0^+} h(x) = 0$, $\lim_{x \nearrow +\infty} h^2(x)(x-\mu_0)p_{\mu_0}(x) = 0$ and the expectations are finite,

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

Example 6. Consider univariate truncated Gaussian distributions with known mean parameter μ and unknown variance parameter $\sigma^2 > 0$. Then, given i.i.d. samples $X_1, \ldots, X_n \sim p_{\sigma_0^2}$, the generalized h-score 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 5, $\lim_{x \nearrow +\infty} h^2(x)(x-\mu)^3 p_{\sigma_0^2}(x) = 0$, then $\sqrt{n}(\hat{\sigma}_h^2 - \sigma_0^2) \rightarrow_d \mathcal{N}(0, \tau^2)$ with

$$\tau^2 \equiv \frac{2\sigma_0^6\mathbb{E}_0[h^2(X)(X-\mu)^2] + \sigma_0^8\mathbb{E}_0[h'^2(X)(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}.$$

When $\mu_0 = 0$, $h(x) \equiv 1$ also satisfies (A1)–(A2) and (C1)–(C2), and the resulting estimator corresponds to the sample variance, which obtains the Cramér-Rao lower bound.

Remark 7. In the case of univariate truncated Gaussian, an intuitive explanation that using a bounded h gives better results than Hyvärinen (2007) goes as follows. When $\mu \gg \sigma$, there is effectively no truncation to the Gaussian distribution, and our method automatically adapts to using low moments in (4), since a bounded and increasing $h(x)$ becomes almost constant as it gets close to its asymptote for x large. When $h(x)$ becomes constant, we get back to the original score matching for distributions on $\mathbb{R}$. In other cases, the truncation effect is significant, and similar to Hyvärinen (2007), our estimator uses higher moments accordingly.

Figure 1 shows the theoretical asymptotic variance of $\hat{\mu}_h$ as given in Example 5, with $\sigma = 1$ known. Efficiency curves measured by the Cramér-Rao lower bound divided by the asymptotic variance are 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 small finite samples.

Figure 2 is an analog of Figure 1 for $\hat{\sigma}_h^2$ assuming $\mu = 0.5$. For demonstration we choose a nonzero μ since when $\mu = 0$ one can simply use the sample variance (degree of freedom unadjusted) which achieves the Cramér-Rao bound.

Here, the truncated versions of x and x^2 have similar performance when σ is not too small. In fact, when σ is small, the truncation effect is small and one does not lose much by using the sample variance.

4 REGULARIZED GENERALIZED SCORE MATCHING

We now turn to high-dimensional problems, where the number of parameters r is larger than the sample size n. Targeting sparsity in the parameter $\boldsymbol{\theta}$, we consider ℓ_1 regularization (Tibshirani, 1996), which has also been used in graphical model estimation (Meinshausen and Bhlmann, 2006; Yuan and Lin, 2007; Voorman et al., 2013).

Definition 8. Assume the data matrix $\mathbf{x} \in \mathbb{R}^{n \times m}$ comprises n i.i.d. samples from distribution P_0 with
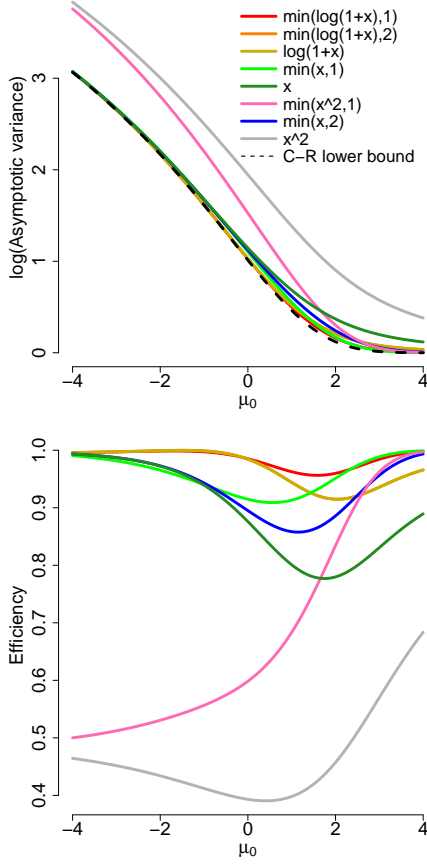
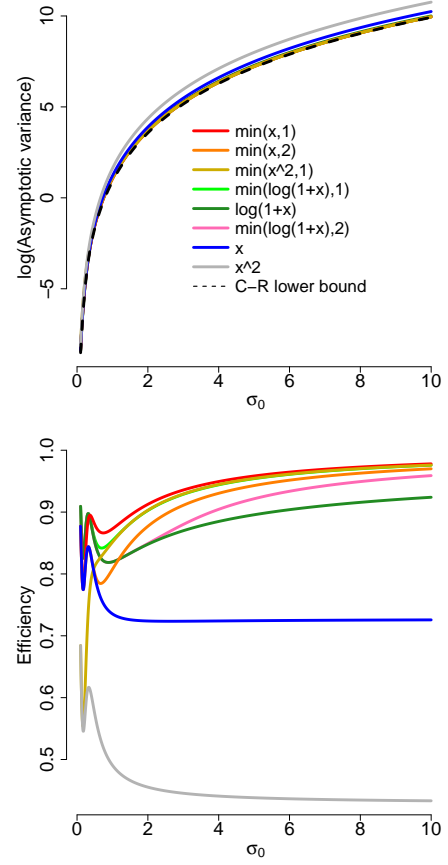Figure 1: Log of asymptotic variance and efficiency w.r.t. the Cramér-Rao bound for $\hat{\mu}_h$ ($\sigma^2 = 1$ known).

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

density p_0 that belongs to an exponential family $\{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \subset \mathcal{P}_+$, where $\boldsymbol{\Theta} \in \mathbb{R}^r$. Define the *regularized generalized h-score matching estimator* as

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \hat{J}_{\boldsymbol{h},r}(\boldsymbol{\theta}) \tag{7}$$

$$\equiv \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x}) \boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1,$$

where $\lambda \geq 0$ is a tuning parameter, and $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are from (6).

Discussion of the general conditions for almost sure uniqueness of the solution is omitted here, but in practice we multiply the diagonals of $\boldsymbol{\Gamma}$ by a constant slightly larger than 1 that ensures strict convexity and thus uniqueness of solution. Details about estimation consistency after this operation will be presented in future work. When the solution is unique, the solution path is *piecewise linear*; compare Lin et al. (2016).

4.1 Truncated GGMs

Using notation from the introduction, let $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathbf{K})$ be a truncated normal random vector with

mean parameter $\boldsymbol{\mu}$ and inverse covariance/precision matrix parameter $\mathbf{K}$. Recall that the conditional independence graph for $\boldsymbol{X}$ corresponds to the support of $\mathbf{K}$, defined as $S \equiv S(\mathbf{K}) \equiv \{(i,j) : \kappa_{ij} \neq 0\}$. This support is our target of estimation.

4.2 Truncated Centered GGMs

Consider the case where the mean parameter is zero, i.e., $\boldsymbol{\mu} \equiv \mathbf{0}$, and we want to estimate the inverse covariance matrix $\mathbf{K} \in \mathbb{R}^{m^2}$. Assume that for all j there exist constants M_j and M_j' that bound h_j and its derivative h_j' a.e. Assume further that $h_j(x) > 0$ a.e, $\lim_{x \searrow 0^+} h_j(x) = 0$, and $h_j'(x) \geq 0$. Boundedness here is for ease of proof in the main theorems; reasonable choices of unbounded h are also valid. Then, (A1)–(A2) are satisfied, and the loss can be written as

$$\hat{J}_{r+}(\mathbf{K}) = \frac{1}{2} \mathrm{vec}(\mathbf{K})^\top \boldsymbol{\Gamma}(\mathbf{x}) \mathrm{vec}(\mathbf{K})$$
$$- \boldsymbol{g}(\mathbf{x})^\top \mathrm{vec}(\mathbf{K}) + \lambda \|\mathbf{K}\|_1, \tag{8}$$

with the jth block of the $m^2 \times m^2$ block-diagonal matrix $\boldsymbol{\Gamma}(\mathbf{x})$ being

$$n^{-1}\mathbf{x}^\top \text{diag}(\boldsymbol{h}_j(\boldsymbol{X}_j))\mathbf{x},$$

where $\boldsymbol{h}_j(\boldsymbol{X}_j) \equiv [h_j(X_j^{(1)}), \ldots, h_j(X_j^{(n)})]^\top$, $\text{diag}(c_1, \ldots, c_n)$ denotes a diagonal matrix with diagonal entries $c_1, \ldots, c_n$, and

$$\boldsymbol{g}(\mathbf{x}) \equiv \text{vec}(\mathbf{u}) + \text{vec}(\text{diag}(\boldsymbol{V})), \quad \mathbf{u} \equiv n^{-1}\mathbf{h}'(\mathbf{x})^\top \mathbf{x},$$

$$\boldsymbol{V} \equiv n^{-1}\mathbf{h}(\mathbf{x})^\top \mathbf{1}_n, \qquad \mathbf{h}(\mathbf{x}) \equiv [h_j(X_j^{(i)})]_{i,j},$$

$$\mathbf{h}'(\mathbf{x}) \equiv [h_j'(X_j^{(i)})]_{i,j}, \qquad \mathbf{1}_n = [1, \ldots, 1]^\top.$$

The *regularized generalized $\boldsymbol{h}$-score matching estimator* of $\mathbf{K}$ in the truncated centered GGM is

$$\hat{\mathbf{K}} \equiv \underset{\mathbf{K} \in \mathbb{R}^{m^2}, \mathbf{K} = \mathbf{K}^\top}{\text{argmin}} \hat{J}_{r+}(\mathbf{K}), \tag{9}$$

where $\boldsymbol{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}(\mathbf{x})$ are defined above.

Definition 9. For true inverse covariance matrix $\mathbf{K}_0$, let $\boldsymbol{\Gamma}_0 \equiv \mathbb{E}_0 \boldsymbol{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}_0 \equiv \mathbb{E}_0 \boldsymbol{g}(\mathbf{x})$. Denote the support of a precision matrix $\mathbf{K}$ as $S \equiv S(\mathbf{K}) \equiv \{(i,j) : \kappa_{ij} \neq 0\}$. Write the true support of $\mathbf{K}_0$ as $S_0 = S(\mathbf{K}_0)$. Suppose the maximum number of non-zero entries in rows of $\mathbf{K}_0$ is $d_{\mathbf{K}_0}$. Let $\boldsymbol{\Gamma}_{SS}$ be the entries of $\boldsymbol{\Gamma} \in \mathbb{R}^{m^2 \times m^2}$ corresponding to edges in S. Define

$$c_{\boldsymbol{\Gamma}_0} \equiv \|(\boldsymbol{\Gamma}_{0,S_0 S_0})^{-1}\|_{\infty,\infty}, \quad c_{\boldsymbol{K}_0} \equiv \|\mathbf{K}_0\|_{\infty,\infty}.$$

We say the irrepresentability condition holds for $\boldsymbol{\Gamma}_0$ if there exists an $\alpha \in (0,1]$ such that

$$\|\boldsymbol{\Gamma}_{0,S_0^c S_0}(\boldsymbol{\Gamma}_{0,S_0 S_0})^{-1}\|_{\infty,\infty} \leq (1 - \alpha). \tag{10}$$

Theorem 10. *Suppose $\boldsymbol{X} \sim \text{TN}(\mathbf{0}, \mathbf{K}_0)$ and $\boldsymbol{h}$ is as discussed in the opening paragraph of this section. Suppose further that $\boldsymbol{\Gamma}_{0,S_0 S_0}$ is invertible and satisfies the irrepresentability condition (10) with $\alpha \in (0,1]$. Let $\tau > 3$. If the sample size and the regularization parameter satisfy*

$$n \geq \mathcal{O}\left(d_{\mathbf{K}_0}^2 \tau \log m \max\left\{\frac{c_{\boldsymbol{\Gamma}_0}^2 c_{\boldsymbol{X}}^4}{\alpha^2}, 1\right\}\right), \tag{11}$$

$$\lambda > \mathcal{O}\left[(c_{\mathbf{K}_0} c_{\boldsymbol{X}}^2 + c_{\boldsymbol{X}} + 1)\left(\sqrt{\frac{\tau \log m}{n}} + \frac{\tau \log m}{n}\right)\right], \tag{12}$$

where $c_{\boldsymbol{X}} \equiv 2\max_j\left(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\mathbb{E}_0 X_j\right)$, then the following statements hold with probability $1 - m^{3-\tau}$:

(a) The regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\mathbf{K}}$ defined in (9) is unique, has its support

included in the true support, $\hat{S} \equiv S(\hat{\mathbf{K}}) \subseteq S_0$, and

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

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

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

(b) Moreover, if

$$\min_{j,k}|\kappa_{0,jk}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2 - \alpha}\lambda,$$

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

The theorem is proved in the supplement. A key ingredient of the proof is a tail bound on $\|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_0\|_\infty$, which is composed of products of the $X_j^{(i)}$'s. In Lin et al. (2016), the products are up to fourth moments. Using a bounded $\boldsymbol{h}$ our products automatically calibrate to a quadratic polynomial when the observed values are large, and resort to higher moments only when they are small. Using this scaling, we obtain improved bounds and convergence rates, underscored in the new requirement on the sample size n, which should be compared to $n > \mathcal{O}((\log m^\tau)^8)$ in Lin et al. (2016).

4.3 Truncated Non-centered GGMs

Suppose now $\boldsymbol{X} \sim \text{TN}(\boldsymbol{\mu}_0, \mathbf{K}_0)$ with both $\boldsymbol{\mu}_0$ and $\mathbf{K}_0$ unknown. While our main focus is still on the inverse covariance parameter $\mathbf{K}$, we now also have to estimate a mean parameter $\boldsymbol{\mu}$. Instead, we estimate the canonical parameters $\mathbf{K}$ and $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$. Concatenating $\boldsymbol{\Xi} \equiv [\mathbf{K}, \boldsymbol{\eta}]$, the corresponding $\boldsymbol{h}$-score matching loss has a similar form to (8) and (9) with $\mathbf{K}$ replaced by $\boldsymbol{\Xi}$, and different $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$. As a corollary of the centered case, we have an analogous bound on the error in the resulting estimator $\hat{\boldsymbol{\Xi}}$; we omit the details here. We note, however, that we can have different tuning penalty parameters $\lambda_{\mathbf{K}}$ and $\lambda_{\boldsymbol{\eta}}$ for $\mathbf{K}$ and $\boldsymbol{\eta}$, respectively, as long as their ratio is fixed, since we can scale the $\boldsymbol{\eta}$ parameter by the ratio accordingly. To avoid picking two tuning parameters, one may also choose to remove the penalty on $\boldsymbol{\eta}$ altogether by profiling out $\boldsymbol{\eta}$. We leave a detailed analysis of the profiled estimator to future research.

4.4 Tuning Parameter Selection

By treating the loss as the mean negative log-likelihood, we may use the extended Bayesian information Criterion (eBIC) to choose the tuning parameter (Chen and Chen, 2008; Foygel and Drton, 2010). Let

$\hat{S}^\lambda \equiv \{(i,j) : \hat{\kappa}_{ij}^\lambda \neq 0, i < j\}$, where $\hat{\mathbf{K}}^\lambda$ be the estimate associated with tuning parameter λ. The eBIC is then

$$\text{eBIC}(\lambda) = -n\text{vec}(\hat{\mathbf{K}})^\top \boldsymbol{\Gamma}(\mathbf{x})\text{vec}(\hat{\mathbf{K}})$$

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

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

5 NUMERICAL EXPERIMENTS

We present simulation results for non-negative GGM estimators with different choices of $\boldsymbol{h}$. We use a common h for all columns of the data matrix $\mathbf{x}$. Specifically, we consider functions such as $h_j(x) = x$ and $h_j(x) = \log(1+x)$ as well as truncations of these functions. In addition, we try MCP (Fan and Li, 2001) and SCAD penalty-like (Zhang, 2010) functions.

5.1 Implementation

We use a coordinate-descent method analogous to Algorithm 2 in Lin et al. (2016), where in each step we update each element of $\hat{\mathbf{K}}$ based on the other entries from the previous steps, while maintaining symmetry. Warm starts using the solution from the previous λ, as well as lasso-type strong screening rules (Tibshirani et al., 2012) are used for speedups. In our simulations below we always scaled the data matrix by column ℓ_2 norms before proceeding to estimation.

5.2 Truncated Centered GGMs

For data from a truncated centered Gaussian distribution, we compared our truncated centered Gaussian estimator (9) with various choices of h, to *SpaCE JAM* (SJ, Voorman et al., 2013), which estimates graphs without assuming a specific form of distribution, a pseudo-likelihood method *SPACE* (Peng et al., 2009) with *CONCORD* reformulation (Khare et al., 2015), *graphical lasso* (Yuan and Lin, 2007; Friedman et al., 2008), the *neighborhood selection* estimator (NS) in Meinshausen and Bühlmann (2006), and finally *nonparanormal SKEPTIC* (Liu et al., 2012) with Kendall's τ. While we ran all these competitors, only the top performing methods are explicitly shown in our reported results. Recall that the choice of $h(x) = x^2$ corresponds to the estimator in Lin et al. (2016), using score matching from Hyvärinen (2007).

A few representative ROC (*receiver operating characteristic*) curves for edge recovery are plotted in Figures
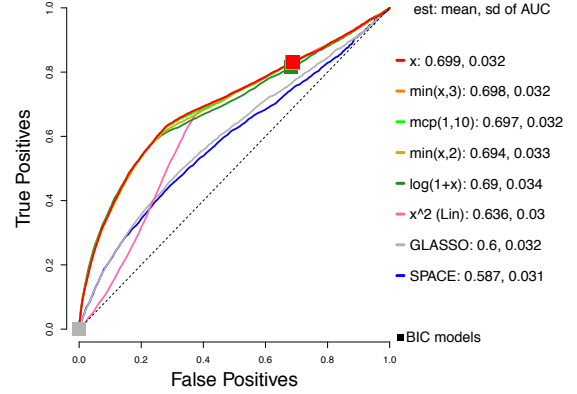

Figure 3: Average ROC curves of our estimator with various choices of h, compared to SPACE and GLASSO, for the *truncated centered Gaussian* case. $n = 80$, $m = 100$. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC (with refitting) for the estimator in the same color. eBIC is introduced in Section 4.4.
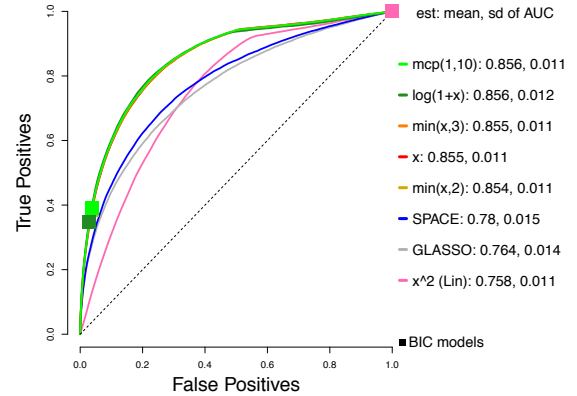

Figure 4: Same setting as in Figure 3, with $n = 1000$.

3 and 4, using $m = 100$ and $n = 80$ or $n = 1000$, respectively. Each curve corresponds to the average of 50 ROCs obtained from estimation of $\mathbf{K}$ from $\mathbf{x}$ generated using 5 different true precision matrices $\mathbf{K}_0$, each with 10 trials. The averaging method is mean AUC-preserving and is introduced as *vertical averaging* and outlined in Algorithm 3 in Fawcett (2006). The construction of $\mathbf{K}_0$ is the same as in Section 4.2 of Lin et al. (2016): a graph with $m = 100$ nodes with 10 disconnected subgraphs containing the same number of nodes, i.e. $\mathbf{K}_0$ is block-diagonal. In each sub-matrix, we generate each lower triangular element to be 0 with probability $\pi \in (0, 1)$, and from a uniform distribution on interval $[0.5, 1]$ with probability $1 - \pi$. The upper triangular elements are set accordingly by symmetry. The diagonal elements of $\mathbf{K}_0$ are chosen to be a common positive value so that the minimum eigenvalue of
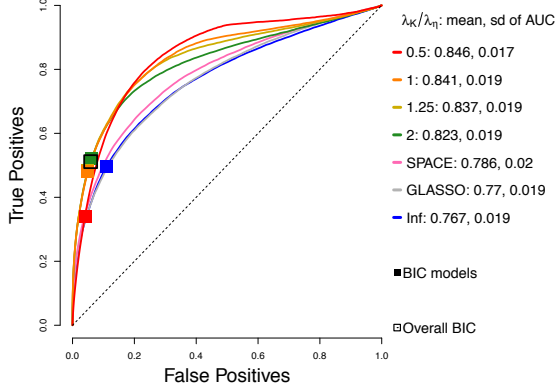
Figure 5: Performance of the non-centered estimator with $h(x) = \min(x, 3)$; each curve corresponds to a different choice of fixed $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$. $n = 1000$, $m = 100$.

Figure 6: Performance of the non-centered profiled estimators using various h, $n = 1000$, $m = 100$.

$\mathbf{K}_0$ is 0.1. We choose $\pi = 0.2$ for $n = 80$ and $\pi = 0.8$ for $n = 1000$.

For clarity, we only plot some top-performing representatives of the functions we considered. However, all of the alternative functions h we considered perform better than $h(x) = x^2$ from Hyvärinen (2007) and Lin et al. (2016).

5.3 Truncated Non-Centered GGMs

Next we generate data from a truncated non-centered Gaussian distribution with both parameters $\boldsymbol{\mu}$ and $\mathbf{K}$ unknown. Consider $\hat{\mathbf{K}}$ as part of the estimated $\hat{\boldsymbol{\Xi}}$ as discussed in Section 4.3. In each trial we form the true $\mathbf{K}_0$ as in Section 5.2, and we generate each component of $\boldsymbol{\mu}_0$ independently from the normal distribution with mean 0 and standard deviation 0.5.

As discussed in Section 4.3, we assume the ratio of the tuning parameters for $\mathbf{K}$ and $\boldsymbol{\eta}$ to be fixed. Shown in Figure 5 are average ROC curves (over 50 trials as in Section 5.2) for truncated non-centered GGM estimators with $h(x) = \min(x, 3)$; each curve corresponds to a different ratio $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$, where "Inf" indicates $\lambda_{\boldsymbol{\eta}} \equiv 0$. Here, $n = 1000$ and $m = 100$.

Clearly, as the ratio increases, the performance improves, and after a certain threshold it deteriorates. The AUC for the profiled estimator with $\lambda_{\boldsymbol{\eta}} = 0$ is among the worst, so there indeed is a lot to be gained from tuning an extra tuning parameter, although there is a tradeoff between time and performance.

In Figure 6 we compare the performance of the profiled estimator with different h, to SPACE and GLASSO, each with 50 trials as before. It can be seen that even without tuning the extra parameter, the estimators, except for $h(x) = x^2$, still work as well as SPACE and
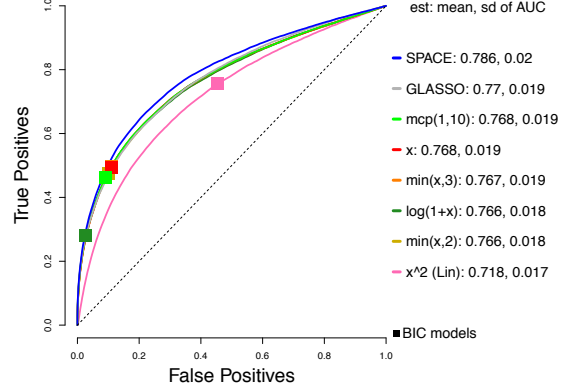
GLASSO, and outperform Lin et al. (2016).

6 DISCUSSION

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

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

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

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

References

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

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

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

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

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

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

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

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

B. Fellinghauer, P. Bühlmann, M. Ryffel, M. Von Rhein, and J. D. Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.

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

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

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

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

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

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

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

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

H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, et al. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

N. Meinshausen and P. Bhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

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

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

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3): 1287–1319, 2010.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.

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

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16 (1):3813–3847, 2015.

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