Electronic Journal of Statistics
Vol. 10 (2016) 806–854
ISSN: 1935-7524
DOI: 10.1214/16-EJS1126

Estimation of high-dimensional graphical models using regularized score matching

Lina Lin, Mathias Drton

Department of Statistics, University of Washington, Seattle, WA 98195, USA
e-mail: linlina@uw.edu; md5@uw.edu

and

Ali Shojaie

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
e-mail: ashojaie@uw.edu

Abstract: Graphical models are widely used to model stochastic dependences among large collections of variables. We introduce a new method of estimating undirected conditional independence graphs based on the score matching loss, introduced by Hyvärinen (2005), and subsequently extended in Hyvärinen (2007). The *regularized score matching* method we propose applies to settings with continuous observations and allows for computationally efficient treatment of possibly non-Gaussian exponential family models. In the well-explored Gaussian setting, regularized score matching avoids issues of asymmetry that arise when applying the technique of neighborhood selection, and compared to existing methods that directly yield symmetric estimates, the score matching approach has the advantage that the considered loss is quadratic and gives piecewise linear solution paths under ℓ_1 regularization. Under suitable irrepresentability conditions, we show that ℓ_1-regularized score matching is consistent for graph estimation in sparse high-dimensional settings. Through numerical experiments and an application to RNAseq data, we confirm that regularized score matching achieves state-of-the-art performance in the Gaussian case and provides a valuable tool for computationally efficient estimation in non-Gaussian graphical models.

MSC 2010 subject classifications: Primary 62H12; secondary 62F12.
Keywords and phrases: Conditional independence graph, exponential family, graphical model, high-dimensional statistics, score matching, sparsity.

Received September 2015.

1. Introduction

Undirected graphical models, also known as *Markov random fields*, are important tools for summarizing dependency relationships between random variables and have found application in many fields, including bioinformatics, language and speech processing, and digital communications. Each such model is associated to

an undirected graph $G = (V, E)$, with vertex set V and edge set $E \subset V \times V$. For a random vector $X = (X_j : j \in V)$ indexed by the nodes of G, the graphical model given by G requires that X_j and X_k be conditionally independent given all other variables whenever nodes j and k are not joined by an edge in G (Lauritzen, 1996). If G is the smallest graph such that X satisfies this requirement, we term G the *conditional independence graph* of X. In this case, X_j and X_k are conditionally independent given all other variables if and only if j and k are non-adjacent in G. We will always take the vertex set to be $V = \{1, \ldots, m\}$, so m is the number of observed variables in X.

Specific models are obtained from additional distributional assumptions. Particularly, an assumption of multivariate normality gives Gaussian graphical models, for which estimation of conditional independence graphs is equivalent to *covariance selection* (Dempster, 1972). If X is jointly multivariate normal with mean vector μ and covariance matrix $\mathbf{\Sigma}$—in symbols, $X \sim N(\mu, \mathbf{\Sigma})$—then the conditional independences among the random variables, and hence edges between nodes in the graph, are determined by the entries of the inverse covariance, or concentration matrix $\mathbf{K} = (\kappa_{jk}) = \mathbf{\Sigma}^{-1}$. More precisely, $\kappa_{jk} = 0$ for $j \neq k$ if and only if X_j and X_k are independent given all other variables.

There is a large literature on selection of conditional independence graphs; see the references in Edwards (2000, Chap. 6) or Drton and Perlman (2007). In the last decade, attention has shifted to high-dimensional settings with the number of variables m comparable to or larger than the sample size n. This scenario arises, for instance, in microarray experiments. Fortunately, high-dimensional problems may remain tractable in the presence of structural constraints such as *sparsity*, i.e., if each node in the graph is incident to a small number of edges. This is of interest for microarray data as gene regulatory networks are intrinsically sparse (Leclerc, 2008).

Gaussian models have been the primary tool for graphical modeling of data comprising continuous variables, such as gene expression data, and a large number of methods have been proposed for statistical estimation in high-dimensional Gaussian graphical models. A common strategy involves augmenting a loss function with a sparsity-inducing penalty such as an ℓ_1, or lasso penalty. Two widely-used approaches are the *graphical lasso* or *glasso* (Yuan and Lin, 2007) and *neighborhood selection* (Meinshausen and Bühlmann, 2006). In glasso, an ℓ_1 penalty on the entries of the inverse covariance matrix is added to the negative Gaussian log-likelihood. Neighborhood selection, on the other hand, is an ℓ_1-penalized pseudo-likelihood approach that leverages the fact that the node-wise full conditional distributions from a Gaussian graphical model form m linear regression models. Meinshausen and Bühlmann (2006) treat these separate regression models as having their parameters unrelated, but as we discuss below, methods that account for the symmetry in a concentration matrix have been proposed in subsequent work.

Methods for high-dimensional data have also been developed for non-Gaussian settings. Miyamura and Kano (2006), Finegold and Drton (2011), Vogel and Fried (2011) and Sun and Li (2012) address robustness to outliers. Liu, Lafferty and Wasserman (2009), Liu et al. (2012) and Dobra and Lenkoski (2011) treat

Gaussian copula models. Neighborhood selection/pseudo-likelihood procedures can also be applied to models for categorical models where the node-wise regression is logistic or multinomial (Lee, Ganapathi and Koller, 2007; Höfling and Tibshirani, 2009; Ravikumar, Wainwright and Lafferty, 2010; Jalali et al., 2011). Allen and Liu (2013) and Yang et al. (2012) discuss extensions using node-wise generalized linear models, and semi-/nonparametric methods were proposed by Fellinghauer et al. (2013) and Voorman, Shojaie and Witten (2014).

In this paper, we propose a different approach to high-dimensional graphical model selection. Addressing the case of *continuous* but not necessarily Gaussian observations, the proposed method is based on the *score matching* loss, first introduced by Hyvärinen (2005) in the setting of image analysis. Recently, Forbes and Lauritzen (2015) studied score matching in Gaussian graphical models with symmetry constraints, and demonstrated that, when the number of variables m is fixed, the estimators derived from the score matching loss are asymptotically efficient in some special cases, but not in general. Our focus is instead on the use of score matching in high-dimensional problems, for which we consider regularization with an ℓ_1 penalty. We will refer to this graphical model selection technique as *regularized score matching*.

Regularized score matching is computationally very convenient for any exponential family comprising continuous distributions. Indeed, the score matching loss is a positive semi-definite quadratic function. It follows that the solution path for the regularized score matching problem is piecewise linear and can be computed in its entirety. Moreover, theoretical analysis can be based on familiar techniques. Most importantly, as we demonstrate for Gaussian graphical models, regularized score matching exhibits state-of-the-art statistical efficiency in high-dimensional settings. The method also performs well in our applications to non-Gaussian models, which include models that seem rather difficult to handle via other methods.

In the Gaussian setting, regularized score matching is structurally closest to pseudo-likelihood methods with symmetry constraints, such as *SPACE* (Peng et al., 2009), *symmetric lasso* (Friedman, Hastie and Tibshirani, 2010) and *SPLICE* (Rocha, Zhao and Yu, 2008). A thorough discussion of these different methods is given by Khare, Oh and Rajaratnam (2015) who also reformulate the SPACE objective function to ensure convergence of coordinate descent algorithms. They abbreviate their method as *CONCORD*. For brevity, we refer to these algorithms collectively as SPACE. We note that in contrast to regularized score matching, the SPACE methods do not have piecewise linear solution paths. Furthermore, as remarked before, the computational convenience of regularized score matching carries over to non-Gaussian settings.

A limitation of the original score matching introduced by Hyvärinen (2005) is that it requires the data to be generated from a distribution whose density is twice differentiable on $\mathbb{R}^m$. Hyvärinen (2007) proposed a generalization of the approach to the important case of non-negative data. For exponential families, the non-negative score matching loss is again a semidefinite quadratic function. We explore regularization of the non-negative score matching loss as a tool for estimation of conditional independence graphs from high-dimensional non-

negative data, and we establish consistency of the method.

The remainder of the paper is organized as follows. Section 2 provides the needed background on score matching and its applications. In Section 3, we describe the proposed method, *regularized* score matching. Implementation details are given in Appendix A. In Section 4, we present results of numerical experiments to compare the performance of the procedure with existing approaches. An application to RNAseq data is given in Section 5. Section 6 provides sparsistency theory for both basic and non-negative regularized score matching. Proofs are given in Section 7 with details deferred to Appendix B and C. We end with a discussion in Section 8. Computer code used in our numerical work is provided as supplementary material (Lin, Drton and Shojaie, 2016).

Notation

The following notational conventions are used throughout the paper:

(i) Random variables/vectors are denoted by upper case letters; lower case letters are used for observed values. So, $x \in \mathbb{R}^m$ is an observed value of the random vector X. Similarly, $\mathbf{x} = (x_{ij}) \in \mathbb{R}^{n\times m}$ is a matrix of observed values, which will typically hold the realizations of n i.i.d. copies of X in its rows. We index the columns of a matrix with subscripts, so x_j refers to the jth column of $\mathbf{x}$. Superscripts in parentheses are used to refer to the rows of a matrix, so $x^{(i)}$ is the ith row of $\mathbf{x}$.

(ii) For a matrix $\mathbf{U} = (u_{ij}) \in \mathbb{R}^{m\times m}$, we denote the vectorization obtained by stacking columns by

$$\operatorname{vec}(\mathbf{U}) = \left(u_{11}, u_{21}, \ldots, u_{m1}, \ldots, u_{1m}, \ldots, u_{mm}\right)^T.$$

(iii) Let $a, b \in [1, \infty]$. We denote the ℓ_a norm of a vector $u \in \mathbb{R}^m$ by

$$\|u\|_a \;=\; \left(\sum_{i=1}^{m} |u_i|^a\right)^{1/a}$$

and write $|||\mathbf{U}|||_{a,b} = \max_{\|\boldsymbol{x}\|_a=1} \|\mathbf{U}\mathbf{x}\|_b$ for the ℓ_a/ℓ_b operator norm of a matrix $\mathbf{U} \in \mathbb{R}^{m\times m}$. We let $|||\mathbf{U}|||_\infty = |||\mathbf{U}|||_{\infty,\infty}$ and $\|\mathbf{U}\|_a = \|\operatorname{vec}(\mathbf{U})\|_a$.

2. Score matching

We begin with an overview of Hyvärinen's score matching, discussing first random vectors supported on all of $\mathbb{R}^m$ and then random vectors supported on the nonnegative orthant. We also review the convenient form of the score matching estimating equations in exponential families.

2.1. Basic score matching

Suppose X is a continuous random vector taking values in $\mathbb{R}^m$, with joint distribution P. Suppose further that P belongs to the family $\mathcal{P}$ that comprises all

probability distributions with support equal to $\mathbb{R}^m$ and a twice differentiable density with respect to Lebesgue measure. We emphasize that in a statistical context the differentiability requirement is with respect to data. We write p to denote the density of P and adopt the usual notation for the gradient and Laplacian

$$\nabla f(x) = \left\{ \frac{\partial}{\partial x_j} f(x) \right\} \in \mathbb{R}^m, \qquad \Delta f(x) = \sum_{j=1}^{m} \frac{\partial^2}{\partial x_j^2} f(x) \in \mathbb{R},$$

of a function $f : \mathbb{R}^m \to \mathbb{R}$.

For a distribution $Q \in \mathcal{P}$ with density q, define the divergence function

$$J(Q) = \int_{\mathbb{R}^m} p(x) \left[\|\nabla \log q(x) - \nabla \log p(x)\|_2^2 \right] dx \tag{2.1}$$

as the expected squared distance between the gradients of the log-densities of the two distributions Q and P. By choosing Q to minimize (2.1), we are matching 'scores' with respect to the data vector x. Hence, (2.1) has been referred to as the *score matching loss*. It is evident from (2.1) that the score matching loss is uniquely minimized when $Q = P$.

Upon initial inspection, optimization of $J(Q)$ seems to require knowledge of P in an important way. However, Hyvärinen (2005) showed that, under mild regularity conditions, the score matching loss (2.1) can be rewritten as:

$$J(Q) = \int_{\mathbb{R}^m} p(x) \left[\Delta \log q(x) + \frac{1}{2} \|\nabla \log q(x)\|_2^2 \right] dx \; + \; \text{const}, \tag{2.2}$$

where 'const' refers to a term independent of Q. The key term in the integrand in (2.2) is the so-called Hyvärinen scoring rule

$$S(x, Q) = \Delta \log q(x) + \frac{1}{2} \|\nabla \log q(x)\|_2^2.$$

The integral in (2.2) admits an empirical version in which the integration with respect to P is replaced by an average over an observed sample, which we arrange into a data matrix $\mathbf{x} \in \mathbb{R}^{n \times m}$. This leads to the *empirical score matching loss*

$$\hat{J}(\mathbf{x}, Q) = \frac{1}{n} \sum_{i=1}^{n} S(x^{(i)}, Q), \tag{2.3}$$

and the *score matching estimator* (SME)

$$\hat{Q} = \arg\min_{Q} \hat{J}(\mathbf{x}, Q).$$

The score matching loss $J(Q)$ was motivated by problems involving models whose distributions have an intractable normalization constant. Indeed, evaluating (2.2) and computing the SME $\hat{Q}$ requires no knowledge of the normalization constant, which is eliminated upon taking logarithmic derivatives with respect to

x. Besides the imaging problems considered by Hyvärinen (2005), score matching has been applied to spatial statistics (Dawid and Musio, 2013) and neural networks (Köster and Hyvärinen, 2007; Vincent, 2011; Le et al., 2011).

The statistical properties of SMEs in classical large sample settings have been investigated by Hyvärinen (2005, 2007) and Forbes and Lauritzen (2015). In particular, it has been shown that, under the usual regularity conditions, SMEs are asymptotically consistent and normal in large-sample theory. However, SMEs are not necessarily asymptotically efficient.

2.2. Extension to non-negative data

The partial integration arguments underlying (2.2) may fail to apply when considering distributions Q that are not supported on all of $\mathbb{R}^m$. In particular, when Q is taken to be from $\mathcal{P}_+$, i.e. the family of distributions that are supported on $\mathbb{R}^m_+ = [0,\infty)^m$ with Lebesgue densities that are twice differentiable on $(0,\infty)^m$, then partial integration may not be possible due to discontinuities at points with zero coordinates. We thus consider the non-negative score matching loss,

$$J_+(Q) = \int_{\mathbb{R}^m_+} p(x) \left[\left\| \nabla \log q(x) \circ x - \nabla \log p(x) \circ x \right\|_2^2 \right] dx, \tag{2.4}$$

as proposed in Hyvärinen (2007). Here, '$\circ$' stands for the Hadamard product, that is, element-wise multiplication.

The score matching loss (2.1) can be thought of as a function of the Euclidean distance between the gradients of the model density q and true density p with respect to a hypothetical location parameter μ, evaluated at 0. That is, we may write (2.1) as

$$J(Q) = \int_{\mathbb{R}^m} p(\mathbf{x}) \left[\left\| \nabla_{\mu=0} \log q(x+\mu) - \nabla_{\mu=0} \log p(x+\mu) \right\|_2^2 \right] dx.$$

Likewise, the non-negative score matching loss compares the gradient of the model density q and true density p with respect to a hypothetical scale parameter σ evaluated at 1,

$$J_+(Q) = \int_{\mathbb{R}^m_+} p(\mathbf{x}) \left[\left\| \nabla_{\sigma=1} \log q(x \circ \sigma) - \nabla_{\sigma=1} \log p(x \circ \sigma) \right\|_2^2 \right] dx.$$

Under suitably adjusted regularity conditions, Hyvärinen (2007) showed that the non-negative score matching loss from (2.4) can be simplified into

$$J_+(Q) = \int_{\mathbb{R}^m_+} p(x) S_+(x,Q)\, dx \;+\; \text{const} \tag{2.5}$$

with scoring rule

$$S_+(x,Q) = \sum_{j=1}^m \left[2x_j \frac{\partial \log q(x)}{\partial x_j} + x_j^2 \frac{\partial^2 \log q(x)}{\partial x_j^2} + \frac{1}{2} x_j^2 \left(\frac{\partial \log q(x)}{\partial x_j} \right)^2 \right]. \tag{2.6}$$

For a data matrix $\mathbf{x} \in \mathbb{R}^{n\times m}$, one obtains the *empirical non-negative score matching loss*

$$\hat{J}_+(\mathbf{x}, Q) = \frac{1}{n}\sum_{i=1}^{n} S_+(x^{(i)}, Q), \tag{2.7}$$

and the *non-negative score matching estimator* (SME_+)

$$\hat{Q}_+ = \arg\min_Q \hat{J}_+(\mathbf{x}, Q).$$

Again, under the usual regularity conditions, the estimator $\hat{Q}_+$ is asymptotically consistent and normal in traditional large-sample theory.

2.3. Score matching in exponential families

Hyvärinen (2007) and Forbes and Lauritzen (2015) have shown that the SME has a convenient closed form as a rational function of the data when $\mathcal{P}$ is an exponential family. Hyvärinen (2007) showed the same for SME_+ for the example of truncated normal distributions. As they provide the basis for our later work, we revisit these results for both SME and SME_+.

Let $\mathcal{P} = (Q_\theta : \theta \in \Theta)$ be an exponential family with natural parameter space Θ. Suppose that the distributions Q_θ have their common support equal to either $\mathcal{X} = \mathbb{R}^m$ or $\mathcal{X} = \mathbb{R}^m_+$, and that $\mathcal{P}$ is dominated by Lebesgue measure on $\mathbb{R}^m$. Assuming that the sufficient statistics $t(x)$ take values in $\mathbb{R}^s$, the log-densities of the distributions Q_θ have the form

$$\log q(x|\theta) = \theta^T t(x) - \psi(\theta) + b(x), \quad x \in \mathcal{X}, \tag{2.8}$$

and

$$\Theta = \left\{\theta \in \mathbb{R}^s \ : \ \psi(\theta) = \log \int_{\mathcal{X}} e^{\theta^T t(x)} dx < \infty\right\}. \tag{2.9}$$

Lemma 1. *Let $\mathbf{x} \in \mathbb{R}^{n\times m}$ be a data matrix, and suppose $\mathcal{P} = (Q_\theta : \theta \in \Theta)$ is an exponential family characterized by (2.8) and (2.9). If $\mathcal{P}$ has support $\mathcal{X} = \mathbb{R}^m$, then the empirical score matching loss $\hat{J}(\mathbf{x}, Q_\theta)$ is a quadratic function in θ with*

$$\hat{J}(\mathbf{x}, Q_\theta) = \frac{1}{2}\,\theta^T \mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x})^T\theta + c(\mathbf{x}), \tag{2.10}$$

where $\mathbf{\Gamma}(\mathbf{x})$ is a positive semidefinite $s \times s$ matrix, and $g(\mathbf{x})$ is an s-vector. The same is true for $\hat{J}_+(\mathbf{x}, Q_\theta)$ when $\mathcal{P}$ has support $\mathcal{X} = \mathbb{R}^m_+$.

Proof. For $j = 1, \ldots, m$ and $x \in \mathbb{R}^m$, define the s-vectors

$$h_j(x) = \frac{\partial}{\partial x_j} t(x), \qquad h_{jj}(x) = \frac{\partial^2}{\partial x_j^2} t(x).$$

It then follows from (2.8) that $\hat{J}(\mathbf{x}, Q_\theta)$ can be expressed in the claimed form with

$$\mathbf{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(x^{(i)})h_j(x^{(i)})^T, \tag{2.11}$$

$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\partial}{\partial x_j}b(x^{(i)})\right)h_j(x^{(i)})^T + \Delta t(x^{(i)}), \tag{2.12}$$

$$c(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\|\nabla b(x^{(i)})\right\|_2^2 + \Delta b(x^{(i)}). \tag{2.13}$$

For non-negative score matching, $\hat{J}_+(\mathbf{x}, Q_\theta)$ admits the claimed form with

$$\mathbf{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2 h_j(x^{(i)})h_j(x^{(i)})^T, \tag{2.14}$$

$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\partial}{\partial x_j}b(x^{(i)})\right)h_j(x^{(i)})^T + x_{ij}^2 h_{jj}(x^{(i)})^T + 2x_j^{(i)}h_j(x^{(i)})^T, \tag{2.15}$$

$$c(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{2}x_{ij}^2\left(\frac{\partial}{\partial x_j}b(x^{(i)})\right)^2 + x_{ij}^2\frac{\partial^2}{\partial x_j^2}b(x^{(i)}) + 2x_{ij}\frac{\partial}{\partial x_j}b(x^{(i)}), \tag{2.16}$$

where the x_{ij} are the entries of the $n \times m$ data matrix $\mathbf{x}$. □

Lemma 1 implies that, when working with exponential families, both score matching objectives are quadratic functions of the unknown parameter vector θ. A score matching estimator $\hat{\theta}$ thus satisfies a set of *linear* estimating equations

$$\hat{\theta}^T\mathbf{\Gamma}(\mathbf{x}) + g(\mathbf{x}) = 0. \tag{2.17}$$

2.4. Pairwise interaction models

The most basic class of exponential families that appear in graphical modeling are pairwise interaction models with log-densities

$$\log q(x|\theta) \;=\; \sum_{1\le j\le k\le m}\theta_{jk}t_{jk}(x_j, x_k) - \psi(\theta) + b(x), \quad x \in \mathcal{X} \subseteq \mathbb{R}^m. \tag{2.18}$$

Here, the t_{jk} are sufficient statistics that depend only on the jth and kth coordinate of x, and the θ_{jk} are interaction parameters. If Q_θ denotes the distribution with density given by (2.18), then the Hammersley-Clifford Theorem implies that an edge between nodes j and k exists in the conditional independence graph of Q_θ if and only if θ_{jk} is nonzero. The specific models we consider later

either exactly have the form in (2.18) or are closely related extensions with log-densities

$$\log q(x|\theta) \;=\; \sum_{a=1}^{A}\sum_{j\le k}\theta_{jk}^{(a)}t_{jk}^{(a)}(x_j,x_k)+\sum_{l=1}^{L}\sum_{j=1}^{m}\theta_j^{(l)}t_j^{(l)}(x_j)-\psi(\theta)+b(x), \quad (2.19)$$

where pairwise interactions may be of A different types and we also include L sets of sufficient statistics $t_j^{(l)}$ depending on the individual coordinates. The latter appear, for instance, when allowing distributions to vary in location. The distribution Q_θ defined by (2.19) has no edge between j and k in its conditional independence graph if and only if $\theta_{jk}^{(1)}=\cdots=\theta_{jk}^{(A)}=0$.

In our study of score matching methods for models of the type (2.18) or (2.19), it will be convenient to introduce the symmetric $m\times m$ interaction matrix $\mathbf{\Theta}$ with entries

$$\mathbf{\Theta}_{jk}=\begin{cases}\theta_{jk} & \text{if } \; j\le k,\\ \theta_{kj} & \text{if } \; j>k.\end{cases}$$

Lemma 2. *Let $\mathcal{P}$ to be the pairwise interaction model given by (2.18) with symmetric $m\times m$ interaction matrix $\mathbf{\Theta}$. If $\mathcal{P}$ has support $\mathcal{X}=\mathbb{R}^m$, then the empirical score matching loss $\hat{J}(\mathbf{x},Q_\theta)$ equals*

$$\frac{1}{2}vec(\mathbf{\Theta})^T\mathbf{\Gamma}(\mathbf{x})vec(\mathbf{\Theta})+g(\mathbf{x})^T vec(\mathbf{\Theta})+c(\mathbf{x}) \quad (2.20)$$

for a symmetric $m^2\times m^2$ matrix $\mathbf{\Gamma}(\mathbf{x})$ that is block-diagonal, with all blocks of size $m\times m$. The same is true for $\hat{J}_+(\mathbf{x},Q_\theta)$ when $\mathcal{P}$ has support $\mathcal{X}=\mathbb{R}_+^m$.

Proof. By (2.11) and (2.14), it suffices to show that there exists a block-diagonal matrix $\mathbf{\Gamma}_j(x)$ such that

$$\theta^T h_j(x)h_j(x)^T\theta \;=\; \mathrm{vec}(\mathbf{\Theta})^T\mathbf{\Gamma}_j(x)\mathrm{vec}(\mathbf{\Theta}), \quad (2.21)$$

where $\theta=(\theta_{jk}: j\le k)$. Now,

$$\begin{aligned}h_j(x)^T\theta &= \sum_{k\ge j}\frac{\partial}{\partial x_j}t_{jk}(x_j,x_k)\theta_{jk}+\sum_{k<j}\frac{\partial}{\partial x_j}t_{kj}(x_k,x_j)\theta_{kj}\\ &= \sum_{k\ge j}\frac{\partial}{\partial x_j}t_{jk}(x_j,x_k)\mathbf{\Theta}_{kj}+\sum_{k<j}\frac{\partial}{\partial x_j}t_{kj}(x_k,x_j)\mathbf{\Theta}_{kj}.\end{aligned}$$

Define a vector $\bar{h}_j(x)\in\mathbb{R}^{m^2}$, indexed by pairs (k,l) with $1\le k,l\le m$, by setting the entries to

$$\bar{h}_j(x)_{kl}=\begin{cases}\frac{\partial}{\partial x_k}t_{kl}(x_k,x_l) & \text{if } \; j=k\le l,\\ \frac{\partial}{\partial x_k}t_{lk}(x_k,x_j) & \text{if } \; j=k>l,\\ 0 & \text{if } \; j\ne k.\end{cases} \quad (2.22)$$

Then $h_j(x)^T\theta=\bar{h}_j(x)\mathrm{vec}(\mathbf{\Theta})$ and (2.21) holds with $\mathbf{\Gamma}_j(x)=\bar{h}_j(x)\bar{h}_j(x)^T$, which is block-diagonal as it is zero with the exception of the $m\times m$ block indexed by pairs (k,l) with $k=j$. □

Remark 1. When $\mathcal{P}$ is a model as specified in (2.19), then the empirical (non-negative) score matching loss may still be represented as an explicit quadratic form with a block-diagonal symmetric matrix $\mathbf{\Gamma}(\mathbf{x})$ as in (2.20). However, $\mathbf{\Gamma}(\mathbf{x})$ is then of size $(Am^2 + Lm) \times (Am^2 + Lm)$, and its m diagonal blocks are of size $(Am + L) \times (Am + L)$. The jth block has its rows and columns corresponding to the jth columns of each of $\mathbf{\Theta}^{(1)}, \ldots, \mathbf{\Theta}^{(A)}$ as well $(\theta_j^{(1)}, \ldots, \theta_j^{(L)})$.

Example 1. If the exponential family is taken to be the family of centered multivariate normal distributions with precision matrix $\mathbf{K} = (\kappa_{jk})$, then the support is $\mathcal{X} = \mathbb{R}^m$ and

$$q(x|\mathbf{K}) \propto \exp\left\{-\frac{1}{2}x^T\mathbf{K}x\right\}, \quad x \in \mathbb{R}^m. \tag{2.23}$$

With

$$\nabla \log q(x|\mathbf{K}) = -\mathbf{K}x, \qquad \Delta \log q(x|\mathbf{K}) = -\sum_{j=1}^{m} \kappa_{jj},$$

and dropping a term that is constant in $\mathbf{K}$, the empirical score matching loss from (2.2) takes the form

$$-\ \mathrm{tr}(\mathbf{K}) + \frac{1}{2}\mathrm{tr}(\mathbf{KKW}), \tag{2.24}$$

where

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}x^{(i)T}$$

is the empirical covariance matrix (under knowledge of zero mean). Lemma 2 applies with $t_{jk}(x_j, x_k) = x_j x_k$, in which case the matrix $\mathbf{\Gamma}_j(x)$ constructed in the proof of the lemma does not depend on j, other than through the location of the nonzero block. Indeed, (2.20) holds with $\mathbf{\Gamma}(\mathbf{x}) = \mathbf{I}_{m\times m} \otimes \mathbf{W}$ and $g(\mathbf{x}) = \mathrm{vec}(\mathbf{I}_{m\times m})$, where $\mathbf{I}_{m\times m}$ is the $m \times m$ identity matrix. Clearly, $\mathbf{\Gamma}(\mathbf{x})$ is positive definite if and only if $\mathbf{W}$ is as well. If $\mathbf{W}$ is invertible then SME of $\mathbf{K}$ is $\hat{\mathbf{K}} = \mathbf{W}^{-1}$ and coincides with the maximum likelihood estimator.

Example 2. Consider truncated normal densities of the form

$$q(x|\mathbf{K}) \propto \exp\left\{-\frac{1}{2}x^T\mathbf{K}x\right\}, \quad x \in \mathbb{R}_+^m. \tag{2.25}$$

Using κ_j to denote the jth column of $\mathbf{K}$, it can be shown that the empirical non-negative score matching objective is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} 2x_{ij}x^{(i)T}\kappa_j - x_{ij}^2\kappa_{jj} + \frac{1}{2}\kappa_j^T\left(x_{ij}^2 x^{(i)}x^{(i)T}\right)\kappa_j. \tag{2.26}$$

The loss can be written as in (2.10) with $\mathbf{\Gamma}(\mathbf{x})$ a block diagonal $m^2 \times m^2$ matrix, whose jth block is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 x^{(i)} x^{(i)T}.$$

Moreover, $g(\mathbf{x}) = 2w + w_{\text{diag}}$, where $w = \text{vec}(\mathbf{W})$ and $w_{\text{diag}} = \text{vec}(\text{diag}(\mathbf{W}))$. The maximum likelihood estimator for $\mathbf{K}$ has no closed form due to intractable normalizing constants.

Example 3. Finally, consider the family of distributions with densities of the form

$$q(x|\mathbf{B}^{(2)}, \mathbf{B}, \mathbf{b}) \propto \exp\left\{ \sum_{1\le j\ne k\le m} \beta_{jk}^{(2)} x_j^2 x_k^2 + \sum_{j,k=1}^{m} \beta_{jk} x_j x_k + \sum_{j=1}^{m} \beta_j x_j \right\}, \quad x \in \mathbb{R}^m. \tag{2.27}$$

Here, $\mathbf{b} = (\beta_1, \ldots, \beta_m)^T$ is an m-vector, and $\mathbf{B} = (\beta_{jk})$ and $\mathbf{B}^{(2)} = (\beta_{jk}^{(2)})$ are symmetric $m \times m$ interaction matrices, the latter having a zero diagonal. This family is a class of distributions with normal conditionals, with densities that need not be unimodal (Arnold, Castillo and Sarabia, 1999; Gelman and Meng, 1991). This family is intriguing from the perspective of graphical modeling as, in contrast to the Gaussian case, conditional dependence may also express itself in the variances. For conditional independence of X_j and X_k both β_{jk} and $\beta_{jk}^{(2)}$ need to vanish.

By Remark 1, the empirical score matching loss for the family from (2.27) can be written as a quadratic function with the quadratic term given by block-diagonal matrix $\mathbf{\Gamma}(\mathbf{x})$ of size $(2m^2 + m) \times (2m^2 + m)$. The blocks are of size $(2m+1) \times (2m+1)$, and the jth block has its rows and columns corresponding to the jth columns of $\mathbf{B}$ and $\mathbf{B}^{(2)}$ and the jth entry in $\mathbf{b}$.

3. Regularized score matching

In this section, we propose the use of *regularized score matching* for graphical model selection in the setting of high-dimensional sparse graphical models. We begin by discussing the proposed method and its implementation. Later sections show that, despite the fact that SMEs need not be asymptotically efficient in the sense of traditional large-sample theory, regularized score matching achieves state-of-the-art statistical performance in high-dimensional problems, all the while allowing seemingly complicated non-Gaussian graphical models to be treated in a computationally efficient manner.

3.1. Methodology

Building on the ideas underlying methods such as glasso, neighborhood selection and SPACE, we augment the score matching loss with a sparsity-promoting

penalty. Our focus is on the most basic case of an ℓ_1 penalty but other regularization schemes could be considered instead; see also Example 3 below.

Using the generic representation given in Lemma 1, for an exponential family, the proposed method is based on minimizing the objective

$$\hat{J}^{\lambda}(\theta) = \frac{1}{2}\,\theta^T \mathbf{\Gamma}(\mathbf{x})\theta - g(\mathbf{x})^T\theta + c(\mathbf{x}) + \lambda\|\theta\|_1, \quad \theta \in \mathbb{R}^s, \tag{3.1}$$

where $\mathbf{\Gamma}(\mathbf{x})$ is positive semidefinite and $\lambda \geq 0$ is a tuning parameter that controls the sparsity level. Larger values of λ yield sparser solutions, and $\lambda = 0$ gives the unregularized SME. Since $\mathbf{\Gamma}(\mathbf{x})$ is positive semidefinite, the function $\hat{J}^{\lambda}(\theta)$ is convex but in the settings of interest here $\mathbf{\Gamma}(\mathbf{x})$ will be singular and $\hat{J}^{\lambda}(\theta)$ will not be strictly convex.

The regularized score matching objective from (3.1) is similar to the lasso objective in linear regression (Tibshirani, 1996), where the function to be minimized takes the special form

$$\frac{1}{2}\|y - X\theta\|_2^2 + \|\theta\|_1, \tag{3.2}$$

for a 'response vector' y and a 'design matrix' X. In the applications we have in mind (3.1) cannot be written exactly as in (3.2) because the vector $g(\mathbf{x})$ is generally not in the column span of $\mathbf{\Gamma}(\mathbf{x})$. However, we may adapt existing optimization methods for lasso to solve the regularized score matching problem. Implementation details are given in Appendix A.

If the considered exponential family is supported on $\mathcal{X} = \mathbb{R}^m$ and we use the loss from (2.3), then we call the minimizer of (3.1) the regularized score matching estimator (rSME). If $\mathcal{X} = \mathbb{R}^m_+$ and we use the loss from (2.7), then we abbreviate to rSME$_+$. In specific instances of graphical models, we may apply the ℓ_1 penalty only to those coordinates of θ whose vanishing corresponds to absence of edges in a conditional independence graph. If the subset $\mathcal{E} \subseteq \{1, \ldots, s\}$ holds the relevant coordinates then we use the penalty

$$\|\theta\|_{1,\mathcal{E}} \equiv \sum_{j \in \mathcal{E}} |\theta_j|.$$

Example 1 (cont.). For the (centered) Gaussian case considered in Example 1, the target of estimation is the symmetric precision matrix $\mathbf{K}$. The conditional independence graph corresponds to the pattern of zeros in the off-diagonal entries of $\mathbf{K}$ and the rSME is

$$\hat{\mathbf{K}} = \arg\min_{\mathbf{K} \in \mathrm{Sym}_m} \left\{ -\operatorname{tr}(\mathbf{K}) + \frac{1}{2}\operatorname{tr}(\mathbf{K}\mathbf{K}\mathbf{W}) + \lambda\|\mathbf{K}\|_{1,\mathrm{off}} \right\}, \tag{3.3}$$

where W is the empirical covariance matrix and $\|\mathbf{K}\|_{1,\mathrm{off}} = \|\mathbf{K}\|_{1,\mathcal{E}}$ penalizes only the off-diagonal entries indexed by $\mathcal{E} = \{(j,k) : j \neq k\}$. We emphasize that while in this example the natural parameter space is the positive definite cone, we propose minimizing simply over the entire space of symmetric $m \times m$ matrices, denoted by Sym_m. As our interest is primarily in graph selection, we

do not enforce positive definiteness of $\hat{\mathbf{K}}$, which is in line with methods such as SPACE or neighborhood selection; compare Khare, Oh and Rajaratnam (2015).

We remark that evaluating the function from (3.3) at a nonsymmetric matrix $\mathbf{K}$ as well as its transpose $\mathbf{K}^T$ gives the same value. By convexity, minimizing over all $m \times m$ matrices gives a solution in Sym_m, which then must equal $\hat{\mathbf{K}}$.

Example 2 (cont.). In the truncated normal family from Example 2, the conditional independence graph corresponds again to the zero pattern in the off-diagonal entries of the positive definite interaction matrix $\mathbf{K}$. Proceeding in analogy to the Gaussian case, we define the rSME_+ as the minimizer $\hat{\mathbf{K}}_+$ of the objective given by (2.26) with the penalty $\lambda\|\mathbf{K}\|_{1,\mathrm{off}}$ added on. Again, we ignore the positive definiteness requirement and minimize the penalized non-negative score matching loss with respect to $\mathbf{K} \in \mathrm{Sym}_m$.

Example 3 (cont.). For the family of distributions with normal conditionals from Example 3, we would like a penalty to induce joint sparsity in the two symmetric interaction matrices $\mathbf{B}$ and $\mathbf{B}^{(2)}$, because an edge between nodes j and k is absent from the conditional independence graph if and only both $\mathbf{B}$ and $\mathbf{B}^{(2)}$ have their (j,k) entries zero. For this purpose, it is natural to adopt the group lasso penalty (Yuan and Lin, 2006). The rSME is then obtained by minimizing the empirical score matching loss augmented by the penalty

$$\lambda \sum_{j \neq k} \sqrt{(\beta_{jk})^2 + (\beta^{(2)}_{jk})^2}.$$

Ignoring again any refined constraints from the natural parameter space of the family, we propose minimizing the penalized loss with respect to $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{B}, \mathbf{B}^{(2)} \in \mathrm{Sym}_m$. Since the group lasso is applied with small groups (of size 2), the problem would be suitable for application of exact block-coordinate descent as discussed in Foygel and Drton (2010a).

3.2. Uniqueness of rSME

In the setup from Lemma 1, we may write

$$\boldsymbol{\Gamma}(\mathbf{x}) = \mathbf{H}(\mathbf{x})^T \mathbf{H}(\mathbf{x}) \tag{3.4}$$

for an $nm \times s$ matrix $\boldsymbol{H}(\mathbf{x})$; recall (2.11) and (2.14). Based on the arguments leading to Lemmas 3 and 5 in Tibshirani (2013), the function $\hat{J}^\lambda(\theta)$ from (3.1) has a unique minimizer $\hat{\theta}$ as long as $\lambda > 0$ and the columns of $\boldsymbol{H}(\mathbf{x})$ are in *general position*. To clarify, suppose that $\mathcal{U} \subset \mathbb{R}^{nm}$ is a collection of $|\mathcal{U}| = s$ vectors. Then $\mathcal{U}$ is in general position if for all $k < \min\{nm, s\}$, all choices of vectors $u_1, \ldots, u_{k+1} \in \mathcal{U}$ and signs $\sigma_1, \ldots, \sigma_{k+1} \in \{-1, 1\}$, the affine span of $\sigma_1 u_1, \ldots, \sigma_{k+1} u_{k+1}$ does not contain any vector u or $-u$ for $u \in \mathcal{U} \setminus \{u_1, \ldots, u_{k+1}\}$.

The graphical models we are interested in are pairwise interaction models that have additional special structure in that the matrix $\boldsymbol{\Gamma}(\mathbf{x})$ is block-diagonal

with m blocks of equal size; recall Lemma 2 and Remark 1. Denote the diagonal blocks by $\mathbf{\Gamma}_1(\mathbf{x}), \ldots, \mathbf{\Gamma}_m(\mathbf{x})$, which in the setup from (2.19) are of size $(Am^2 + Lm) \times (Am^2 + Lm)$. Each block is the sum of n symmetric rank one matrices and we have the decomposition

$$\mathbf{\Gamma}_j(\mathbf{x}) = \mathbf{H}_j(\mathbf{x})^T \mathbf{H}_j(\mathbf{x}), \quad j = 1, \ldots, m. \tag{3.5}$$

The n columns of each of the matrices $\mathbf{H}_j(\mathbf{x})$ were specified in (2.22). It now holds that the regularized score matching problem from (3.1) has a unique minimizer provided each one of the $n \times (Am + L)$ blocks $\mathbf{H}_1(\mathbf{x}), \ldots, \mathbf{H}_m(\mathbf{x})$ defined in (3.5) has its columns in general position.

Example 1 (cont.). In the Gaussian case, $\mathbf{H}_1(\mathbf{x}) = \cdots = \mathbf{H}_m(\mathbf{x}) = \mathbf{x}$. By the Lemma in Okamoto (1973), the set of matrices $\mathbf{x}$ that fail to be in general position has measure zero. The rSME $\hat{\mathbf{K}}$ is unique almost surely when data are generated from a continuous joint distribution.

Example 2 (cont.). In the truncated normal case, $\mathbf{H}_j(\mathbf{x})$ is equal to the matrix obtained from $\mathbf{x}$ by multiplying each column element-wise with x_j, the jth column of $\mathbf{x}$. The Lemma in Okamoto (1973) implies that the rSME$_+$ is unique almost surely.

For the normal conditionals model from Example 3, almost sure uniqueness would have to be derived by appealing to results on uniqueness of group lasso (Roth and Fischer, 2008).

3.3. Piecewise linear paths

The rSME depends on the regularization parameter λ. In this section we make this explicit and denote it by $\hat{\theta}^\lambda$. Adopting standard language, we refer to the set of $\hat{\theta}^\lambda$ obtained by varying λ as the *solution path* and call this path *piecewise linear* if there exists $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_R = \infty$ and $\xi_0, \ldots, \xi_{R-1} \in \mathbb{R}^m$ such that $\hat{\theta}^\lambda = \hat{\theta}^{\lambda_r} + (\lambda - \lambda_r)\xi_r$ for $\lambda \in [\lambda_r, \lambda_{r+1}]$. Piecewise linear solution paths have the appeal that the entire solution path can be found by calculating the change points λ_r and associated slopes ξ_r.

The next lemma is a consequence of the quadratic nature of the score matching objective for exponential families, and holds for the lasso problem as well.

Lemma 3. *The solution path $\hat{\theta}^\lambda$ for the regularized score matching problem from (3.1) is piecewise linear.*

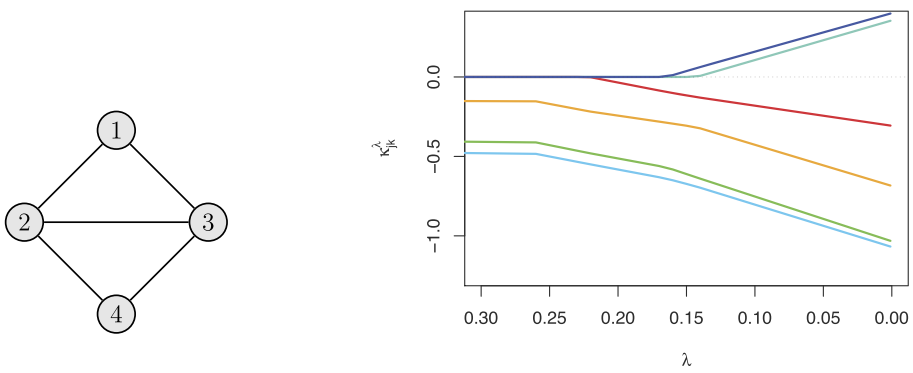
Proof. An s-vector z belongs to $\partial\|\theta\|_1$, the subdifferential of the ℓ_1 norm, if

$$z_j = \begin{cases} \operatorname{sign}(\theta_j) & \text{if } \theta_j \neq 0, \\ \in [-1, 1] & \text{if } \theta_j = 0. \end{cases} \tag{3.6}$$

The Karush-Kuhn-Tucker (KKT) conditions characterizing optimality in (3.1) are

$$\mathbf{\Gamma}(\mathbf{x})\hat{\theta} - g(\mathbf{x}) + \lambda\hat{z} = 0, \quad \hat{z} \in \partial\|\hat{\theta}\|_1. \tag{3.7}$$

The linear relationship between $\hat{\theta}$ and λ (for "fixed" $\hat{z}$) implies the claim. □



(a) (b)

FIG 1. *(a) A conditional independence graph with $m = 4$ nodes. (b) rSME solution path for Gaussian graphical modeling ($m = 4$, $n = 12$).*

While straightforward to show, the property of piecewise linear paths is special to the score matching method we propose. Other methods that give symmetric estimates of precision matrices in Gaussian graphical models, such as glasso or the SPACE-type methods discussed in Khare, Oh and Rajaratnam (2015) do not have piecewise linear solution paths. This said, piecewise linear paths also arise in neighborhood selection (Meinshausen and Bühlmann, 2006), which, however, is a formulation without symmetry. Note also that when using a group lasso penalty as suggested for Example 3, rSME solution paths are no longer piecewise linear.

Example 1 (cont.). In the Gaussian model, the KKT conditions state that $\hat{\mathbf{K}}$ is a solution to (3.1) if and only if

$$(\mathbf{I}_{m\times m} \otimes \mathbf{W})\,\mathrm{vec}(\hat{\mathbf{K}}) - \mathrm{vec}\,(\mathbf{I}_{m\times m}) + \lambda \hat{z} = 0 \tag{3.8}$$

for $\hat{z} \in \partial \|\hat{\mathbf{K}}\|_{1,\mathrm{off}}$, which in slight abuse of notation, we take to mean that

$$\hat{z}_{jk} = \begin{cases} 0 & \text{if } j = k, \\ \mathrm{sign}(\hat{\kappa}_{jk}) & \text{if } \hat{\kappa}_{jk} \neq 0 \text{ and } j \neq k, \\ \in [-1, 1] & \text{if } \hat{\kappa}_{jk} = 0 \text{ and } j \neq k. \end{cases} \tag{3.9}$$

The first case accounts for the fact that the objective is smooth in the diagonal entries of the precision matrix, which are not penalized. Combining (3.8) and (3.9), we have that

$$-1 + \sum_{k=1}^{m} w_{jk}\hat{\kappa}_{jk} = 0, \quad j = 1, \ldots, m, \tag{3.10}$$

$$\sum_{\ell=1}^{m} w_{j\ell}\hat{\kappa}_{\ell k} + \sum_{\ell=1}^{m} w_{k\ell}\hat{\kappa}_{\ell j} + \lambda \hat{z}_{jk} = 0, \quad 1 \leq j \neq k \leq m. \tag{3.11}$$

A Gaussian solution path is shown in Figure 1b, with the horizontal axis transformed to $t(\lambda) = \sum_{j \neq k} |\hat{\kappa}_{jk}^{\lambda}|$. The data were drawn from a multivariate normal distribution with the conditional independence graph from Figure 1a, with sample size $n = 12$. We note that, as one would hope, the coefficient that last enters the solution corresponds to the absent edge $(1, 4)$.

3.4. Tuning

A number of methods have been proposed for selecting the regularization parameter λ in ℓ_1 penalization methods and can be applied in our context. On the one hand, a predictive assessment as in cross-validation can be considered, but the selected graphs are typically too dense. Other possibilities include generalized cross validation (GCV) (Tibshirani, 1996), Akaike's Information Criterion (AIC), approaches based on stability under resampling (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013; Liu, Roeder and Wasserman, 2010), the Bayesian Information Criterion (BIC) (Schwarz, 1978) as well as extensions of BIC proposed to cope with large model spaces (Chen and Chen, 2008; Gao et al., 2012; Foygel and Drton, 2010b; Barber and Drton, 2015). The latter come with some consistency guarantees.

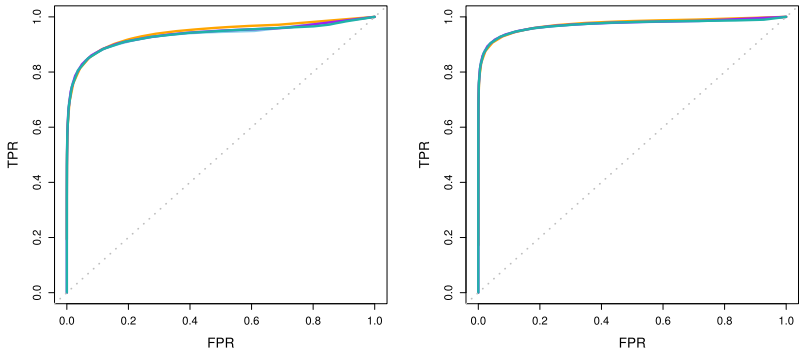
As a demonstration, for the Gaussian case from Example 1, we may consider an extended BIC criterion based on the basic score matching loss (2.2), defined as

$$\mathrm{BIC}(\lambda) = -2\mathrm{tr}(\hat{\mathbf{K}}^{\lambda}) + \mathrm{tr}(\hat{\mathbf{K}}^{\lambda}\hat{\mathbf{K}}^{\lambda}\mathbf{W}) + |\hat{E}^{\lambda}| \log n + 4|\hat{E}^{\lambda}|\gamma \log m, \tag{3.12}$$

where $\hat{E}^{\lambda} = \{(j,k) : \hat{\kappa}_{jk}^{\lambda} \neq 0, j < k\}$ and γ is typically taken to be $1/2$ or 1. Alternatively, we could refit, that is, replace $\mathbf{K}^{\lambda}$ by an unregularized SME computed in the submodel given by constraining all κ_{jk} with $(j,k) \notin \hat{E}^{\lambda}$ to be zero. In either case, we choose the λ which minimizes (3.12).

4. Numerical experiments

We perform numerical experiments comparing regularized score matching to existing methods when data is simulated from (i) a multivariate normal distribution, (ii) a multivariate truncated normal distribution, and (iii) a distribution with normal conditionals. The comparison is made against three methods for estimation of Gaussian graphical models, namely, glasso, neighborhood selection (both implemented in the `R` packages `huge`) and SPACE (in its CONCORD formulation, with `R` package `gconcord`). In addition, we consider the *nonparanormal SKEPTIC*, which applies glasso to a matrix of rank correlations (Kendall's τ or Spearman's ρ) and can be motivated by a Gaussian copula model (Liu et al., 2012). We utilize the version based on Kendall's τ. Finally, we compare to *SPACEJAM* (Voorman, Shojaie and Witten, 2014), which is based on additive modeling of conditional means and implemented in the `R` package `spacejam`. We conclude this section with brief investigations on the robustness of regularized score matching when data is not generated under the assumed model. All



(a) $n = 600$ (b) $n = 1000$

FIG 2. *ROC curves for the Gaussian case. The dashed grey line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), neighborhood selection (——), glasso (——), and SPACE (——). The curves are almost perfectly aligned.*

results in this section are based on averaging over 100 independently generated datasets.

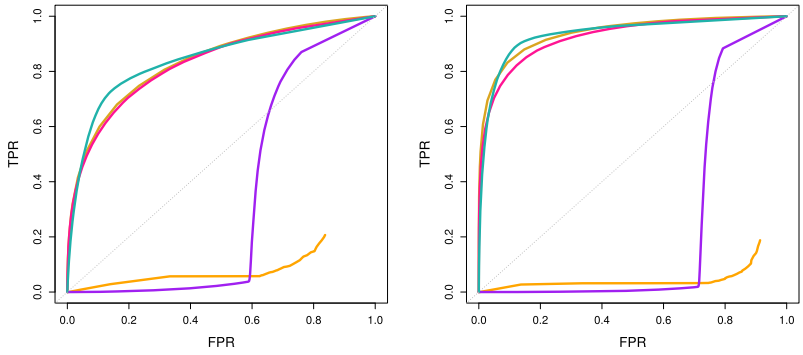
4.1. Gaussian data

We consider a graph with $m = 1000$ nodes, composed of 10 connected components, each 100 nodes in size and structured as a 10×10 2-D lattice (4 nearest neighbors). Each connected component also features three hubs with node degree 20, randomly selected from the subset of nodes in the component.

We follow a procedure similar to the one from Peng et al. (2009) to convert the adjacency matrix of the graph into a sparse diagonally dominant partial correlation matrix. For each non-zero element of the adjacency matrix, we sample a draw from a uniform distribution on $[0.5, 1]$. Each row of this new matrix is then rescaled by 1.5 times the sum of the absolute values of the off-diagonal entries in the row. We average this matrix with its transpose to ensure symmetry, and set its diagonal elements to 1. This matrix is inverted and converted into a correlation matrix to form $\mathbf{\Sigma}^*$.

Data is then generated from a multivariate normal distribution with mean zero and a covariance matrix $\mathbf{\Sigma}^*$. We choose sample size $n = 600$ and 1000. The setup agrees with that in Peng et al. (2009), except that the number of nodes has been scaled up.

Figure 2 shows the ROC curves obtained under both sample sizes. Since the truth is Gaussian, we do not report results for SKEPTIC or SPACEJAM. For both sample sizes, the curve for regularized score matching almost perfectly aligns with those for neighborhood selection, SPACE, and glasso. The results indicate that regularized score matching estimators achieves state-of-the-art statistical efficiency in Gaussian models.



(a) $n = 2500$ (b) $n = 5000$

FIG 3. *ROC curves for the non-negative Gaussian case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), glasso (——), SPACE (——), SKEPTIC (——), and SPACEJAM (——).*

4.2. Non-negative Gaussian data

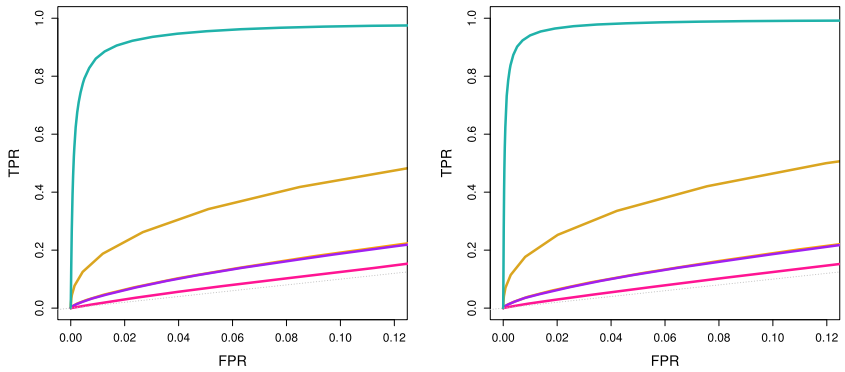
Glasso, SPACE, neighborhood selection and SKEPTIC all presume some form of underlying Gaussianity. In this and the next subsection, we demonstrate the application of regularized score matching in scenarios where these assumptions do not hold to highlight the versatility of the proposed appraoch.

Similar to the Gaussian setting, we consider a graph with $m = 100$ nodes, composed of 10 disconnected subgraphs with equal number of nodes. Using the lower triangular elements adjacency matrix of each 10 node subgraph, we construct ten matrices, where in each matrix, the element is drawn independently to be 0 with probability 0.2, and from a uniform distribution on $[0.5, 1]$ with probability 0.8. The matrices, after symmetrization, are combined into a 100×100 block matrix. The diagonal elements are set to a common positive number such that the minimum eigenvalue is 0.1 to form the precision matrix of the pre-truncated normal, $\mathbf{K}^*$.

Data was then generated from a truncated centered multivariate normal, left-truncated at 0 and with $\mathbf{\Sigma}^* = (\mathbf{K}^*)^{-1}$ as normal covariance. We used the Gibbs sampler from the `tmvtnorm` package in `R` with a burnin period of 100 samples. We thinned out the remaining samples, keeping one in ten. The sample size n is taken to be either 2500 or 5000. The need for a larger sample size is explained by our theoretical findings in Section 6, specifically Corollary 2.

The ROC curves are shown in Figure 3, where regularized score matching outperforms all competitors considered. The closest competitor to regularized score matching are SKEPTIC and SPACEJAM, both of which, objectively, perform well, being capable of capturing some of the non-Gaussianity in the data.

We emphasize that here score matching was applied in its non-negative version from Section 2.2. The basic score matching procedure from Section 2.1 is far less efficient based on experiments not reported here.



(a) $n = 750$ (b) $n = 1000$

FIG 4. *ROC curves for the normal conditionals case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), glasso (——), SPACE (——), SKEPTIC (——), and SPACEJAM (——). The curve for glasso overlaps with the curve for SPACE.*

4.3. Normal conditionals

Next, we take the data-generating distribution to have a density from the class

$$q(x|\mathbf{B}, \mathbf{b}, \mathbf{b}^{(2)}) \propto \exp\left\{\sum_{j\neq k} \beta_{jk} x_j^2 x_k^2 + \sum_{j=1}^{m} \beta_j^{(2)} x_j^2 + \sum_{j=1}^{m} \beta_j x_j\right\}, \quad x \in \mathbb{R}^m, \tag{4.1}$$

where $\mathbf{B} = \{\beta_{jk}\}$ is a symmetric matrix with diagonal entries 0. This family is a special case of the distributions with normal conditionals from Example 3.

We consider the case $m = 625$, with the graph being a 25×25 2-D lattice (4 nearest neighbors). The true interaction matrix $\mathbf{B}^*$ is constructed by multiplying the adjacency matrix by $-1/25$. The coefficients for the terms x_j^2 are all set equal to -1 and those for the x_j all equal to $8/50$, which makes the marginal distributions deviate noticeably from Gaussianity. Data can be generated by Gibbs sampling using the Gaussian full conditionals. We discard the first 100 samples and thin out the remaining samples, keeping one in ten, as in Section 4.2.

We plot the ROC curves for conditional normal data in Figure 4. Regularized score matching outperforms its competitors by a clear margin. This is not surprising, as both glasso and SPACE are derived under normality. A Gaussian copula model as underlying SKEPTIC is of little help. SPACEJAM does best among the competitors but cannot fully extract the available signal about the edge structure as the conditional means are non-additive and the conditional variances are not constant.

4.4. A robustness check

It is of interest to see how score matching performs when the data-generating mechanism is misspecified. We consider two scenarios. First, we apply the Gaussian score matching to a contaminated Gaussian setting similar to that explored in Finegold and Drton (2011). That is, a random subset of Gaussian observations is replaced with Gaussian noise. In the second example, we investigate the performance of the regularized Gaussian score matching when the observations are not Gaussian but rather drawn from a multivariate t-distribution.

4.4.1. Contaminated Gaussians

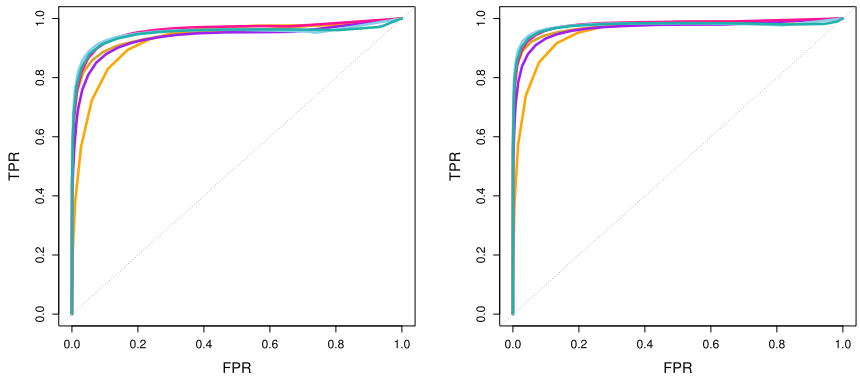
We mimic the setup used in the numerical experiments in Finegold and Drton (2011), who consider these settings to test the robustness of their *tlasso*. Fixing $m = 200$, we construct a sparse precision matrix $\mathbf{K}^*$ according to the following steps: (1) choose each (strictly) lower triangular element of $\mathbf{K}^*$ to be independently -1, 0, 1 with probability 0.01, 0.98 and 0.01 respectively, (2) symmetrize the matrix (3) for each row, i.e. for $j = 1, \ldots, m$, set $\kappa^*_{jj} = 1 + \|\kappa^*_{j,-j}\|_0$ where $\kappa^*_{j,-j}$ refers to the jth row of $\mathbf{K}^*$ with the diagonal element in that row removed. To strengthen partial correlations, the diagonal elements are scaled down by a common positive factor such that the minimum eigenvalue of the resulting matrix is approximately 0.6 (close to 0.62 in our setup). The covariance matrix $\mathbf{\Sigma}^*$ is obtained by inverting $\mathbf{K}^*$.

We generate either $n = 150$ or $n = 200$ observations from a multivariate normal distribution with mean zero and a covariance matrix $\mathbf{\Sigma}^*$. We then corrupt 2% of the observations, substituting them with i.i.d. $N(0, 0.2)$ draws. The corrupted observations cannot easily be differentiated from normal observations, and this elevates the difficulty of the estimation problem.

We present the ROC curves in Figure 5. Interestingly, score matching performs reasonably well, on par with SKEPTIC and neighborhood selection. For both sample sizes, the differences, which are subtle, are most apparent in the regime where the number of false positives detected is small: score matching falls slightly short of neighborhood selection, but it also appears to slightly outperform SKEPTIC. Surprisingly, there is a clear margin of difference between the performances of regularized score matching and SPACE, the former outperforming the latter, despite their noted structural similarities. Glasso, which utilizes the full Gaussian likelihood, performs the worst. Overall, we conclude that regularized score matching is competitively robust when compared to its alternatives in the contaminated Gaussian setting.

4.4.2. Multivariate t-distributed observations

In this section, we apply regularized Gaussian score matching to observations arising from a multivariate t-distribution with mean 0 and covariance matrix $\mathbf{\Sigma}^*$. This corresponds to testing the robustness of regularized score matching



(a) $n = 150$ (b) $n = 200$

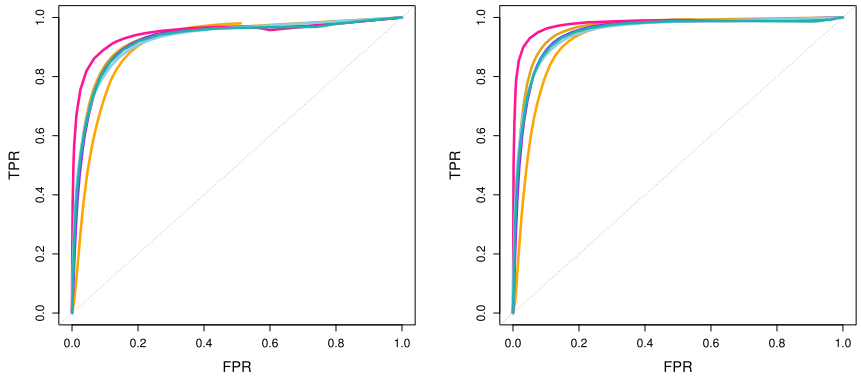
FIG 5. *ROC curves for the contaminated Gaussian case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), neighborhood selection (——), glasso (——), SPACE (——), SKEPTIC (——), and SPACEJAM (——).*

under model misspecification. Like in the previous section, we consider the case when $m = 200$. To set up $\mathbf{\Sigma}^*$, we construct a $m \times m$ adjacency matrix based on an Erdős-Rényi graph with the probability of drawing an edge between any two arbitrary nodes set to 0.01. We then convert the adjacency matrix into $\mathbf{\Sigma}^*$ using the same procedure as in Section 4.1. Samples were drawn from a multivariate t-distribution with covariance matrix $\mathbf{\Sigma}^*$ and three degrees of freedom.

The ROC curves are plotted in Figure 6 for $n = 100$ and $n = 150$. As expected, SKEPTIC outperforms all others, owing to its flexibility to accommodate outliers, as previously demonstrated in Liu et al. (2012). In fact, for elliptical distributions, such as the multivariate t-distribution, Kendall's τ allows for consistent estimation of $\mathbf{\Sigma}^*$, so SKEPTIC should perform optimally (Liu, Han and Zhang, 2012). Nonetheless, regularized score matching is reasonably robust under this setting: its performance is comparable to that of SPACEJAM – only falling slightly short – SPACE, and neighborhood selection. Again, glasso yields the poorest results.

5. Application to RNAseq data

The American Cancer Society estimates that in 2015 there will be 220,800 new cases of prostate cancer and 27,540 deaths. To understand how the cancer develops, as well as how it may be treated, it is necessary to decipher the genetic machinery which drives it. Since cancer is such a complex disease, it is insufficient to study a single gene at a time, as genes may interact with one another in many ways. Graphical modeling of gene expression data has the potential to aid in discovery of such interactions.



(a) $n = 100$ (b) $n = 150$

FIG 6. *ROC curves for the t-distributed case. The dashed line represents random selection of edges. The color to method correspondence is as follows: regularized score matching (——), neighborhood selection (——), glasso (——), SPACE (——), SKEPTIC (——), and SPACEJAM (——).*

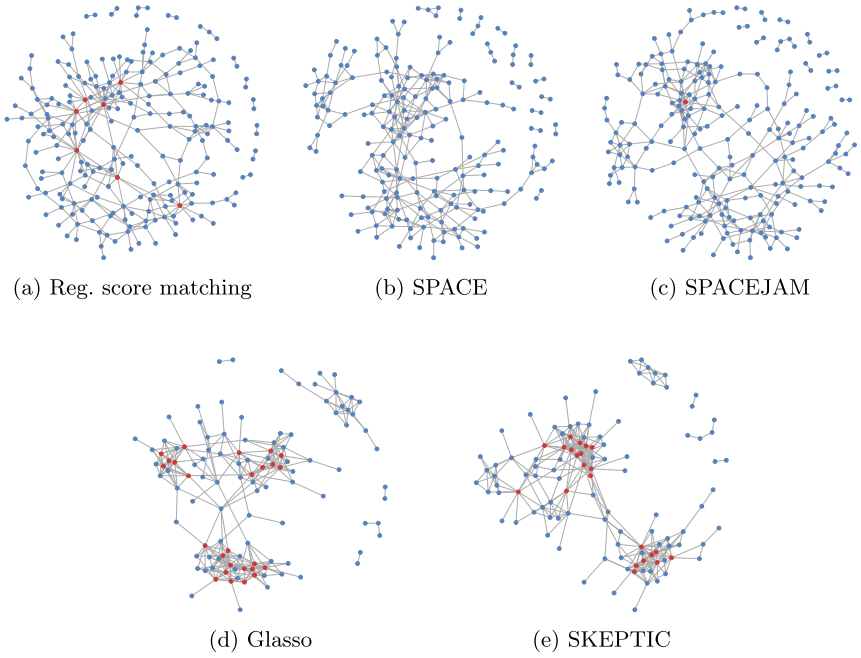
RNAseq data from next-generation sequencing technology can be used to identify genes that are activated/transcribed or suppressed at the time of measurement. However, RNAseq data are non-negative and have skewed marginals, which presents a challenge for existing methodologies. Graphical models based on truncated Gaussian models are interesting alternatives to existing approaches that primarily consist of applying Gaussian methods after transformations. Whether truncation models are truly useful scientifically deserves a fuller exploration; here we simply illustrate how different estimates can be obtained from the proposed methodology.

Our case study is based on the RNAseq data from 487 prostate adenocarcinoma samples available in The Cancer Genome Atlas dataset. We focus on 350 genes that belong to "known" cancer pathways in the Kyoto Encyclopedia of Genes and Genomes. Removing genes with more than 10% missing values, we obtained a dataset with $m = 333$ genes. Remaining missing values were simply set to zero, adding to the challenge. (We will comment on the issue of missing data in the discussion.) In illustration of the regularized score matching methodology, we consider an exponential family of truncated normal distributions with density

$$q(x|\mu, \mathbf{K}) \ \propto \ \exp\left\{\frac{1}{2}(x-\mu)^T\mathbf{K}(x-\mu)\right\}, \quad x \in \mathbb{R}_+^m.$$

This generalizes the family of distributions considered in Example 2 by allowing the truncated normal distribution to have nonzero mean.

We compare regularized non-negative score matching, SPACE (using CONCORD formulation), glasso, SKEPTIC and SPACEJAM. We apply SPACE and glasso directly to the standardized data. We do not consider any marginal trans-



(a) Reg. score matching (b) SPACE (c) SPACEJAM

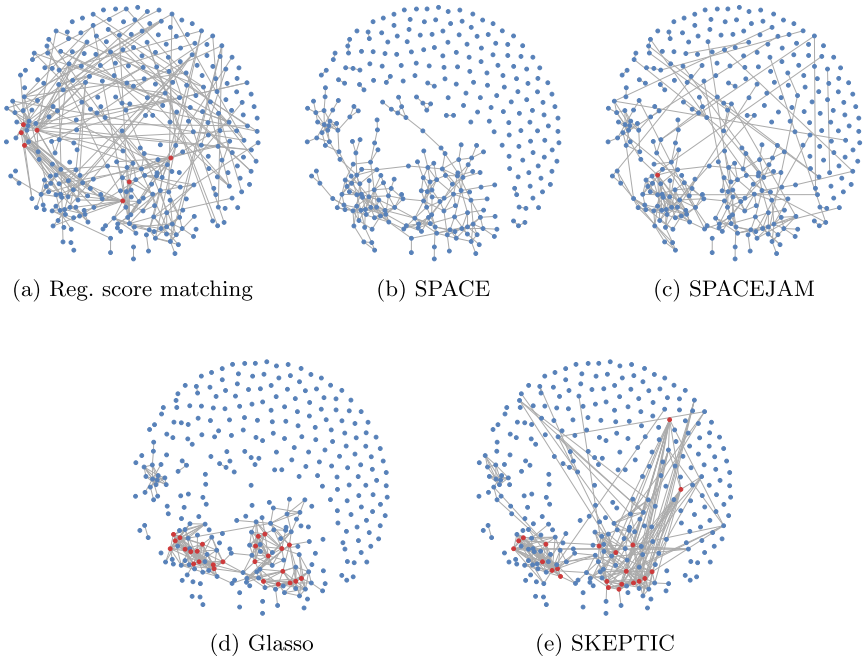
(d) Glasso (e) SKEPTIC

FIG 7. *Topology of inferred networks of* $|E| = 333$ *or* 334 *edges for all considered methods. The layout has been optimized for each graph. Isolated nodes are not shown. Red colored nodes have degree greater or equal to 10.*

formations as they are naturally accounted for when comparing to the rank correlation-based SKEPTIC. For each method, we tune the regularization parameter λ in order to obtain $|E| = 333$ (or 334) edges. Figure 7 depicts the estimated networks, with isolated nodes removed, in layouts optimized for each graph. To allow for easier comparison, we also show the estimated networks in fixed layouts in Figure 8. Node degree distributions are plotted in Figure 9.

By visual inspection, glasso and SKEPTIC give similar topologies, which can be explained by the fact that both are derived from the full Gaussian likelihood. Interestingly, we observe that SPACEJAM and SPACE likewise yield similar graphs, which reinforces findings from Shojaie and Sedaghat (2016). Regularized non-negative score matching yields a graph that is fairly different from the rest.

While the usefulness of these models remains to be further explored, our case study demonstrates that regularized score matching can provide estimates that differ in interesting ways to the estimates generated by other methods. We compile a list of most highly connected genes in each of the estimated graphs in Table 1 (some lists have more than ten genes due to ties), as there is strong evidence that highly connected nodes play important roles in biological networks (Carter et al., 2004; Jeong et al., 2001; Han et al., 2004). There are slight overlaps



(a) Reg. score matching (b) SPACE (c) SPACEJAM

(d) Glasso (e) SKEPTIC

FIG 8. *Topology of inferred networks of* $|E| = 333$ *or* 334 *edges for all considered methods. Layout of nodes is fixed across graph estimates and was optimized for the SPACE estimate. Isolated nodes have now been included. Red colored nodes have degree greater or equal to 10.*

between the lists. Upon further inspection, we observe that six of the ten genes listed under regularized score matching have been previously linked to prostate cancer, five of which have not been identified by the competing methods:

- *CCNE2* (cyclin E2): a protein which is required for transition of the G_1 to S phase of the cell cycle, which determines cell division. Regulated by PTEN, a tumor suppressor, it is over-expressed in metastatic prostate tumor cells (Wu et al., 2009).
- *BRCA2* (breast cancer 2): mutations in the BRCA2 gene have been associated with early-onset prostate cancer in men; men carrying mutations have a predisposition to more aggressive phenotypes (Gayther et al., 2000; Mitra et al., 2008; Tryggvadóttir et al., 2007; Fan et al., 2006).
- *BIRC5* (survivin): a protein which prevents cell death, or apoptosis, and regulates cell division. Heightened expression has been found to be associated with higher final Gleason score, i.e., more aggressive cancer and worse prognosis (Kishi et al., 2004; Shariat et al., 2004).
- *SKP2* (S-phase kinase-associated protein 2, E3 ubiquitin protein ligase): a positive regulator of the G_1 to S phase of the cell cycle, which determines cell division. SKP2 labelling frequency in cancer was positively correlated with the Gleason score, and shown to be a significant predic-

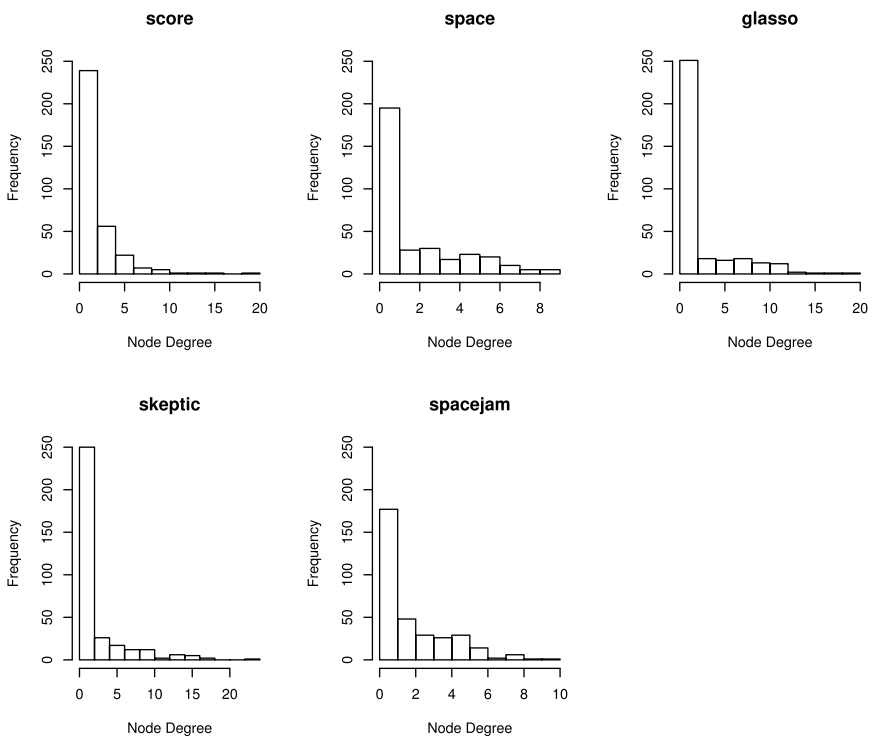


FIG 9. *Node degree distributions for inferred networks of* $|E| = 333$ *or* 334 *edges for all considered methods.*

tor of reduced recurrence-free survival time after radical prostatectomy (Yang et al., 2002; Wang et al., 2008). It has been proposed elsewhere as a promising therapeutic target for prostate cancer (Wang et al., 2012).

- *STAT5B* (signal transducer and activator of transcription 5B): a transcription factor that encourages metastatic behavior of human prostate cancer cells. Its inhibition has been shown to induce apoptosis in human prostate cancer cells (Gu et al., 2010; Ahonen et al., 2003; Moser et al., 2012).

Furthermore, via the Kolmogorov-Smirnov test, we fail to reject the hypothesis that the degrees of the nodes for the regularized score matching graph estimate follow a power law distribution, with significance level of 0.05. On the other hand, we reject this hypothesis for all other generated estimates at the same significance level. There is evidence that genetic networks are 'scale-free', which implies that their degree distribution can be approximated by a power law distribution (Albert, 2005; Barabási and Albert, 1999; Jeong et al., 2001). In this aspect, the topology of regularized score matching estimate is most similar to the hypothesized structure of gene networks.

TABLE 1

The most densely connected genes according to the estimated graphs generated via nonnegative regularized score matching, glasso, SKEPTIC, SPACE and SPACEJAM. The number in parenthesis corresponds to the estimated degree of the gene.

Reg. score matching	Glasso	SKEPTIC	SPACE	SPACEJAM
CCNE2 (19)	EP300 (20)	PIK3CA (23)	TRAF6 (9)	BHX (10)
PIK3CG (16)	SOS1 (17)	FZD7 (18)	TPR (9)	SOS2 (9)
BRCA2 (13)	BAD (16)	PDGFRB (17)	SOS1 (9)	TRAF6 (8)
BIRC5 (12)	TPR (13)	TGFBR2 (16)	JAK1 (9)	TGFBR2 (8)
SKP2 (10)	RBX1 (13)	TCEB2 (16)	EP300 (9)	SOS1 (8)
PIK3CD(10)	PIK3CD (12)	MMP2 (16)	SOS2 (8)	RRM2 (8)
LAMB3 (10)	LAMA4 (12)	LAMA4 (16)	EGFR (8)	PDGFRB (8)
STAT5B (9)	HRAS (12)	GLI2 (15)	CBL (8)	EP300 (8)
HRAS (9)	GLI2 (12)	SOS1 (14)	BAX (8)	PIK3CA (7)
PDGFRB (8)	TRAF6 (11)	PDGFRA (14)	APPL1 (8)	ARNT (7)
GSTP1 (8)	TGFBR2 (11)	MITF (14)		
	TCEB2 (11)	EP300 (14)		
	SPI1 (11)			
	SOS2 (11)			
	PDGFRB (11)			
	MAP2K2 (11)			
	APPL1 (11)			

Finally, we would like to emphasize that we do not intend to claim that regularized score matching provides the *best* estimate of the underlying gene network, as the truth is unknown to us. What we can posit is that truncated Gaussian may be a useful model that provides potentially valid targets for therapy which may be missed by other methods.

6. Theory

This section establishes high-dimensional model selection consistency (sparsistency) of regularized score matching. We focus on pairwise interaction models as in (2.18), although our results could be extended to more general models. Theorem 1 below identifies general deterministic conditions on data that yield sparsistency of regularized (non-negative) score matching. Two subsequent corollaries make probabilistic statements about sparsistency in the Gaussian and the non-negative Gaussian case. Proofs are given in Section 7. Experiments that corroborate the theoretical findings are shown in Appendix C.

Before stating the main results, we describe a key assumption for model selection consistency of ℓ_1-penalized estimators, the irrepresentability assumption, and highlight differences between various estimators of Gaussian graphical models with respect to this assumption.

6.1. Setup and notation

We consider a continuous pairwise interaction model as given by (2.18) with symmetric $m \times m$ interaction matrix $\mathbf{\Theta} = (\theta_{jk})$. We let $\theta = \text{vec}(\mathbf{\Theta})$. Then the

regularized score matching estimator, in its basic or non-negative version, is

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{2}\theta^T \mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x})^T\theta + c(\mathbf{x}) + \lambda\|\theta\|_1. \tag{6.1}$$

By Lemma 2, $\mathbf{\Gamma}(\mathbf{x})$ is a symmetric $m^2 \times m^2$ matrix that is block-diagonal, with blocks of size $m \times m$. For notational convenience, we drop the explicit reference to the data matrix $\mathbf{x}$ and denote $\mathbf{\Gamma}(\mathbf{x})$ and $g(\mathbf{x})$ as $\mathbf{\Gamma}$ and g.

The true data-generating distribution is assumed to belong to the considered model. We denote the true interaction matrix by $\mathbf{\Theta}^* = (\theta^*_{jk})$ and its vectorization by θ^*. We define $\mathbf{\Gamma}^*$ and g^* to be the expected values of $\mathbf{\Gamma}$ and g. The support of θ^*, that is,

$$S \equiv S(\theta^*) = \{(j,k) : j \neq k,\ \theta^*_{jk} \neq 0\}$$

is the edge set of the true conditional independence graph. Similarly,

$$\hat{S} \equiv S(\hat{\theta}) = \{(j,k) : j \neq k,\ \hat{\theta}_{jk} \neq 0\}$$

determines the graph inferred by regularized score matching. Finally, we write d for the maximum degree of the m nodes of the conditional independence graph. In other words, d is the maximum number of nonzero off-diagonal entries in any row (or column) of $\mathbf{\Theta}^*$.

6.2. Irrepresentability

We say that the irrepresentability (or mutual incoherence) condition holds with incoherence parameter α if the following assumption holds.

Assumption 1. *There exists an* $\alpha \in (0,1]$ *such that*

$$\left|\left|\left|\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \leq (1-\alpha). \tag{6.2}$$

Irrepresentability conditions play a key role in the analysis of ℓ_1 regularization techniques (Bühlmann and van de Geer, 2011). For neighborhood selection in Gaussian graphical models, it has been formulated in terms of the covariance matrix $\mathbf{\Sigma}^*$ (Meinshausen and Bühlmann, 2006). In the theoretical analysis of the glasso, the constraint is placed on the Hessian of the log-determinant of the precision matrix $\mathbf{K}^*$, i.e., $(\mathbf{K}^*)^{-1} \otimes (\mathbf{K}^*)^{-1}$ (Ravikumar et al., 2011).

In order to highlight the differences in conditions required for sparsistency of glasso, neighborhood selection, SPACE and regularized score matching, we revisit the Gaussian graphical model example in Meinshausen (2008). Let $\rho \in (0, 1/\sqrt{2})$, and let $\mathbf{\Sigma} = (\sigma_{ij})$ be the 4×4 covariance matrix with ones along the diagonal, $\sigma_{23} = \sigma_{32} = 0$, $\sigma_{14} = \sigma_{41} = 2\rho^2$ and all other off-diagonal entries equal to ρ. The precision matrix $\mathbf{K} = (\mathbf{\Sigma})^{-1}$ then has $\kappa_{14} = \kappa_{41} = 0$. The conditional independence graph G is as in Figure 1a.

Meinshausen showed that for samples drawn from $N(0, \mathbf{\Sigma})$, glasso can consistently recover G only if $\rho \leq \sqrt{3/2} - 1 \approx 0.23$. For neighborhood selection, the

corresponding necessary condition is $\rho \leq 0.5$. If these conditions fail, then for large sample size, the probability of erroneously including the edge $(1, 4)$, i.e., $P(\hat{\kappa}_{14} \neq 0)$ can be shown to be at least 0.5. It turns out that for regularized score matching, the analogous necessary condition gives a bound that falls in between 0.23 and 0.5, specifically, $\rho \leq \sqrt{2} - 1 \approx 0.41$.

We observe that glasso, which yields positive definite estimates, requires the most stringent condition. When working with symmetric matrices as in regularized score matching, the condition is markedly relaxed. Allowing non-symmetric matrices in neighborhood selection leads to further relaxation of the condition. Interestingly, the pseudo-likelihood methods classified under SPACE have the same necessary condition as score matching.

Assumption 1 should be seen as sufficient for consistency of regularized score matching. For Meinshausen's example, it can be shown to amount to $\rho < \frac{1}{2}(\sqrt{3} - 1) \approx 0.37$. The analogous sufficient condition for glasso from Ravikumar et al. (2011) requires that $\rho < \frac{1}{2}(\sqrt{2} - 1) \approx 0.21$. For neighborhood selection, the condition is $\rho < 0.5$.

6.3. Main results

We define

$$c_{\mathbf{\Gamma}^*} = \left|\left|\left|(\mathbf{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty, \text{ and } c_{\mathbf{\Theta}^*} = |||\mathbf{\Theta}^*|||_\infty. \tag{6.3}$$

Moreover, let

$$\mathbf{R}_1 = (\mathbf{\Gamma} - \mathbf{\Gamma}^*), \qquad r_2 = g^* - g, \qquad r_3 = \mathbf{\Gamma}^*\theta^* - g^*, \tag{6.4}$$

such that the KKT conditions from (3.7) can be written as

$$\mathbf{\Gamma}^*(\hat{\theta} - \theta^*) + R_1\hat{\theta} + r_2 + r_3 + \lambda\hat{z} = 0, \quad \hat{z} \in \partial\|\hat{\theta}\|_1. \tag{6.5}$$

Theorem 1. *Assume that $\mathbf{\Gamma}^*_{SS}$ is invertible and the irrepresentability condition holds with incoherence parameter $\alpha \in (0, 1]$ (Assumption 1). Furthermore, assume that*

$$\|\mathbf{R}_1\|_\infty < \epsilon_1, \qquad \|r_2\|_\infty < \epsilon_2, \tag{6.6}$$

with $d\epsilon_1 \leq \alpha/(6c_{\mathbf{\Gamma}^*})$. If

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{\mathbf{\Theta}^*}\epsilon_1, \epsilon_2\}, \tag{6.7}$$

then the following statements hold:

(a) The rSME $\hat{\theta}$ is unique, has its support included in the true support ($\hat{S} \subseteq S$), and satisfies

$$\|\hat{\theta} - \theta^*\|_\infty < \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\lambda.$$

(b) If

$$\min_{1\le j<k\le m} |\theta^*_{jk}| > \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda,$$

then $\hat{S} = S$ *and* $\operatorname{sign}(\hat{\theta}_{jk}) = \operatorname{sign}(\theta^*_{jk})$ *for all* $(j,k) \in S$.

Theorem 1 imposes deterministic conditions on the data, namely, the bounds in (6.6). In the following corollaries, we will consider specific distributional assumptions and impose population conditions that imply bounds of the form (6.6) with high probability.

First, we provide a result for regularized score matching for the Gaussian case (Example 1), which has $\mathbf{\Gamma} = \mathbf{I}_{m\times m} \otimes \mathbf{W}$ with $\mathbf{W}$ being the sample covariance matrix, and $g = \operatorname{vec}(\mathbf{I}_{m\times m})$. When the data is generated from a normal distribution with covariance matrix $\mathbf{\Sigma}^*$ then $\mathbf{\Gamma} = \mathbf{I}_{m\times m} \otimes \mathbf{\Sigma}^*$ and, of course, $g^* = g = \operatorname{vec}(\mathbf{I}_{m\times m})$.

Corollary 1. *Suppose the data is generated from a normal distribution* $N(0, \mathbf{\Sigma}^*)$ *such that* $\mathbf{\Gamma}^*_{SS}$ *is invertible and irrepresentability holds for* $\alpha \in (0,1]$. *Let* $\mathbf{K}^* = (\kappa^*_{jk}) = (\mathbf{\Sigma}^*)^{-1}$,

$$c^* = 3200 \max_{j=1,\dots,m} (\mathbf{\Sigma}^*_{jj})^2 \quad \textit{and} \quad c_1 = \frac{4}{\alpha} c_{\mathbf{\Gamma}^*}.$$

Take any $\tau_1 > 2$. *If the sample size satisfies*

$$n > c^* c_1^2 d^2 (\log m^{\tau_1} + \log 4), \tag{6.8}$$

and the regularization parameter is

$$\lambda > \frac{2c_{\mathbf{K}^*}(2-\alpha)}{\alpha} \sqrt{\frac{c^*(\log m^{\tau_1} + \log 4)}{n}}, \tag{6.9}$$

then the following statements hold with probability $1 - 1/m^{\tau_1 - 2}$:

(a) The rSME $\hat{\mathbf{K}}$ *from (3.3) is unique, has its support included in the true support (*$\hat{S} \subseteq S$*), and satisfies*

$$\|\hat{\mathbf{K}} - \mathbf{K}^*\|_\infty < \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda.$$

(b) If

$$\min_{1\le j<k\le m} |\kappa^*_{jk}| > \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\,\lambda,$$

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

The corollary is proven in Appendix 7.2. Numerical experiments reported in Appendix C suggest that the sample size n indeed needs to scale at least $\Omega(d^2 \log m)$ for sparsistency.

From Theorem 1, we can also derive an analogous result for regularized nonnegative score matching for the truncated Gaussian case (Example 2). The result

requires the sample size to be larger than in the Gaussian case, due to the need to control higher order moments. Recall that here, $\mathbf{\Gamma}(\mathbf{x})$ a block diagonal $m^2 \times m^2$ matrix, with the jth block given by

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 x^{(i)} x^{(i)T},$$

and $g = 2w + w_{\text{diag}}$, where $w = \text{vec}(\mathbf{W})$ and $w_{\text{diag}} = \text{vec}(\text{diag}(\mathbf{W}))$.

Corollary 2. *Suppose the data is generated from a non-negative Gaussian distribution with parameter* $\mathbf{K}^*$*, i.e.,* $N(0, (\mathbf{K}^*)^{-1})$ *is truncated to* $\mathbb{R}_+^m$*. Suppose further that* $\mathbf{\Gamma}^*_{SS}$ *is invertible and irrepresentability holds for* $\alpha \in (0, 1]$*. Let*

$$c^{**} = \max\left\{\left(\frac{L}{2}\right)^4 \sqrt{\max_j \text{Var}[X_j^4]}, \left(\frac{L}{2}\right)^2 \sqrt{\max_j \text{Var}[X_j^2]}\right\} \quad \textit{and} \quad c_2 = \frac{6}{\alpha} c_{\mathbf{\Gamma}^*}$$

where $L > 0$ *is an absolute constant. Take any* $\tau_2 > 3$*. If the sample size satisfies*

$$n > c^{**} c_2^2 d^2 (\log m^{\tau_2} + \log 2)^8, \tag{6.10}$$

and the regularization parameter is

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{\mathbf{K}^*}, 1\} \sqrt{\frac{c^{**}(\log m^{\tau_2} + \log 2)^8}{n}}, \tag{6.11}$$

then the following statements hold with probability $1 - \frac{1}{m^{\tau_2 - 3}}$*:*

(a) The rSME $\hat{\mathbf{K}}_+$ *based on penalizing (2.26) with* $\lambda \|\mathbf{K}\|_{1,off}$ *is unique, has its support included in the true support (*$\hat{S} \subseteq S$*), and satisfies*

$$\|\hat{\mathbf{K}}_+ - \mathbf{K}^*\|_\infty < \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\lambda.$$

(b) If

$$\min_{1 \le j < k \le m} |\kappa^*_{jk}| > \frac{c_{\mathbf{\Gamma}^*}}{2-\alpha}\lambda,$$

then $\hat{S} = S$ *and* $\text{sign}((\hat{\mathbf{K}}_+)_{jk}) = \text{sign}(\kappa^*_{jk})$ *for all* $(j, k) \in S$*.*

The proof of the corollary, which is given in Section 7.3, uses general tail bounds that apply to log-concave measures. The lower bound for n given in (6.10) could well be suboptimal and a lower power of $\log m$ may be sufficient for sparsistency. However, the experiments in Appendix C suggest that the exponent for $\log m$ cannot be taken too much smaller than 8.

We also compared the lower bound we obtained for the non-negative Gaussian case to a result implied by the work of Yang et al. (2013) who treat consistency of neighborhood selection in a general framework that allows node-wise conditional distributions to arise from exponential families. Interestingly, when working out what their general theorem would say about the above non-negative Gaussian model we found that the sample size n would also be required to be at least $\Omega(d^2(\log m)^8)$. Our result from Corollary 2 is thus at least comparable to existing results in the literature.

7. Proofs

7.1. Proof of Theorem 1

First, we note that claim (b) is an immediate consequence of claim (a). To show (a), we apply the primal-dual witness method (PDW) from Wainwright (2009). As explained in detail below, PDW entails construction of a pair $(\tilde{\theta}, \tilde{z})$, with $\tilde{\theta} \in \mathbb{R}^{m^2}$ and $\tilde{z} \in \partial\|\tilde{\theta}\|_1$, that satisfies the KKT optimality conditions from (6.5) and has the support of $\tilde{\theta}$ included in S. If the construction is successful then it ensures that the rSME problem admits a unique solution such that the rSME $\hat{\theta}$ is equal to $\tilde{\theta}$ and inherits all the properties the latter has by definition. These properties include the ℓ_∞ bound on estimation error in addition to the claim about the support.

Replacing $\mathbf{\Gamma}$ by $\mathbf{\Gamma}^*$ and g by g^* in the empirical (basic or non-negative) score matching loss recovers the population loss which, in the present exponential family context, is quadratic and minimized when $\theta = \theta^*$. (Recall that the score matching loss is consistent.) It follows that r_3 from (6.4) is zero as it is the gradient of the population loss. In block form, (6.5) becomes

$$\begin{bmatrix} \mathbf{\Gamma}^*_{SS} & \mathbf{\Gamma}^*_{SS^c} \\ \mathbf{\Gamma}^*_{S^cS} & \mathbf{\Gamma}^*_{S^cS^c} \end{bmatrix} \begin{bmatrix} \hat{\theta}_S - \theta^*_S \\ \hat{\theta}_{S^c} - \theta^*_{S^c} \end{bmatrix} + \begin{bmatrix} \mathbf{R}_{1,SS} & \mathbf{R}_{1,SS^c} \\ \mathbf{R}_{1,S^cS} & \mathbf{R}_{1,S^cS^c} \end{bmatrix} \begin{bmatrix} \hat{\theta}_S \\ \hat{\theta}_{S^c} \end{bmatrix} + \begin{bmatrix} r_{2,S} \\ r_{2,S^c} \end{bmatrix} + \lambda \begin{bmatrix} \hat{z}_S \\ \hat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{7.1}$$

We construct the PDW pair $(\tilde{\theta}, \tilde{z})$ according to the following steps:

(i) Take $\tilde{\theta}$ to be the unique solution to the support-restricted problem, that is,

$$\tilde{\theta} = \arg\min_{\theta_{S^c}=0} \frac{1}{2}\theta^T \mathbf{\Gamma}\theta - g^T\theta + \lambda\|\theta\|_1. \tag{7.2}$$

(ii) Choose

$$\tilde{z}_S \in \partial\|\tilde{\theta}_S\|_1.$$

(iii) Solving (7.1), set

$$\tilde{z}_{S^c} = \frac{1}{\lambda}\Big[- \mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\Big(\mathbf{R}_{1,SS}\tilde{\theta}_S + r_{2,S}\Big) + \mathbf{R}_{1,S^cS}\tilde{\theta}_S + r_{2,S^c} + \lambda\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\tilde{z}_S\Big]. \tag{7.3}$$

(iv) Check the *strict dual feasibility* condition that

$$\|\tilde{z}_{S^c}\|_\infty < 1. \tag{7.4}$$

By step (i), $\tilde{\theta}$ has support contained in S. By step (iii), $(\tilde{\theta}, \tilde{z})$ is guaranteed to fulfill the equations from (7.1). By step (ii), the S-coordinates of $\tilde{z}$ satisfy 'their part' of the subgradient condition. Thus, if the strict dual feasibility from step (iv) holds, then $(\tilde{\theta}, \tilde{z})$ satisfies the KKT conditions from (6.5). Having a strict inequality in (7.4) ensures that every solution to the original rSME problem has support contained in the true support S and since $\mathbf{\Gamma}^*_{SS}$ is assumed invertible,

there is then only one solution (Wainwright, 2009, Lemma 1). The invertibility of $\mathbf{\Gamma}^*_{SS}$ is also what guarantees the uniqueness in step (i).

If the PDW construction is successful, that is, if the strict dual feasibility condition can be established, then we may conclude the rSME $\hat{\theta}$ possesses all the desired properties. Indeed, $\hat{\theta}$ equals $\tilde{\theta}$ which has these properties by construction.

Let $\tilde{\Delta} = \tilde{\theta} - \theta^*$, where $\tilde{\theta}$ is the solution to the support-restricted regularized score matching problem from (7.2). By definition, $\|\tilde{\Delta}\|_\infty = \|\tilde{\Delta}_S\|_\infty$. Furthermore, by step (iii) in the PDW construction,

$$\begin{aligned}\tilde{z}_{S^c} = \frac{1}{\lambda}\Big[\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}(\mathbf{R}_{1,SS}(\theta^*_S + \Delta_S) + r_{2,S}) - \mathbf{R}_{1,S^cS}(\theta^*_S + \Delta_S) - r_{2,S^c}\Big] \\ + \mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\tilde{z}_S. \end{aligned} \tag{7.5}$$

By Assumption 1, and the triangle inequality for the ℓ_∞ norm,

$$\begin{aligned}&\|\tilde{z}_{S^c}\|_\infty \\ &\quad\le \frac{1}{\lambda}\Big[(1-\alpha)\left(\|\mathbf{R}_{1,SS}(\theta^*_S + \Delta_S)\|_\infty + \|r_{2,S}\|_\infty\right) \\ &\qquad + \|\mathbf{R}_{1,S^cS}(\theta^*_S + \Delta_S)\|_\infty + \|r_{2,S^c}\|_\infty\Big] + (1-\alpha) \\ &\quad\le \frac{(2-\alpha)}{\lambda}\Big[\|\mathbf{R}_{1,\cdot S}(\theta^*_S + \Delta_S)\|_\infty + \|r_2\|_\infty\Big] + (1-\alpha) \\ &\quad= \frac{(2-\alpha)}{\lambda}\Big[\|\mathbf{R}_1\theta^* + \mathbf{R}_{1,\cdot S}\Delta_S\|_\infty + \|r_2\|_\infty\Big] + (1-\alpha) \\ &\quad\le \underbrace{\frac{(2-\alpha)}{\lambda}\|\mathbf{R}_1\theta^*\|_\infty}_{=G_1} + \underbrace{\frac{(2-\alpha)}{\lambda}|||\mathbf{R}_{1,\cdot S}|||_\infty\|\Delta_S\|_\infty}_{=G_2} + \underbrace{\frac{(2-\alpha)}{\lambda}\|r_2\|_\infty}_{=G_3} + (1-\alpha),\end{aligned}$$

where the equality in the second to last line follows from the fact that $\theta^*_{S^c} = 0$.

We observe that

$$G_1 = \frac{(2-\alpha)}{\lambda} \times \|\mathbf{\Theta}^*_{\text{wide}}\text{vec}(\mathbf{R}_{1,\text{blocks}})\|_\infty \tag{7.6}$$

where

$$\mathbf{\Theta}^*_{\text{wide}} = \begin{bmatrix} \theta_1^{*T} & 0 & \dots & \dots & 0 \\ 0 & \theta_1^{*T} & 0 & \dots & \vdots \\ \vdots & 0 & \ddots & \ddots & \dots \\ \vdots & \vdots & \ddots & \theta_m^{*T} & 0 \\ \vdots & \vdots & \vdots & 0 & \theta_m^{*T} \end{bmatrix}$$

is an $m^2 \times m^3$ matrix whose diagonal blocks are given by the rows of the the interaction matrix $\mathbf{\Theta}^*$, each row being replicated m times. Moreover, $\text{vec}(\mathbf{R}_{1,\text{blocks}})$ refers to the vectorization of the m diagonal blocks of $\mathbf{R}_1$ that are each of size

$m \times m$; recall Lemma 2. More precisely, if $\mathbf{R}_{1,1}, \ldots, \mathbf{R}_{1,m}$ are the diagonal blocks of $\mathbf{R}_1$, then $\text{vec}(\mathbf{R}_{1,\text{blocks}})$ is obtained by concatenating $\text{vec}(\mathbf{R}_{1,1}), \ldots, \text{vec}(\mathbf{R}_{1,m})$ in that order. Equation (7.6) is the only argument relying on the block-diagonality of $\mathbf{\Gamma}$ and $\mathbf{R}_1$.

From (7.6), we obtain that

$$G_1 \leq \frac{(2-\alpha)}{\lambda} |||\mathbf{\Theta}^*_{\text{wide}}|||_\infty \|\text{vec}(\mathbf{R}_1)\|_\infty < \frac{(2-\alpha)}{\lambda} |||\mathbf{\Theta}^*_{\text{wide}}|||_\infty \epsilon_1.$$

since we have assumed that $\|\text{vec}(\mathbf{R}_1)\|_\infty = \|\mathbf{R}_1\|_\infty < \epsilon_1$. By construction, $|||\mathbf{\Theta}^*_{\text{wide}}|||_\infty = |||\mathbf{\Theta}^*|||_\infty = c_{\mathbf{\Theta}^*}$. It follows, from our choice of λ that $G_1 < \alpha/3$.

By the assumption that $\|r_2\|_\infty < \epsilon_2$, we have

$$G_3 < \frac{(2-\alpha)}{\lambda} \epsilon_2 < \frac{\alpha}{3},$$

and it remains to similarly bound G_2. We treat $|||\mathbf{R}_{1,\cdot S}|||_\infty$ and $\|\tilde{\Delta}_S\|_\infty$ separately.

We note that the rows of $\mathbf{R}_{1,\cdot S}$ have at most d non-zero elements. It follows that $|||\mathbf{R}_{1,\cdot S}|||_\infty \leq d\|\mathbf{R}_1\|_\infty < d\epsilon_1 < \alpha/6c_{\mathbf{\Gamma}}$, where the last inequality holds by assumption. Since $\mathbf{\Gamma}_{SS}$ is assumed invertible, we have from the top block of equations in (7.1) that

$$\tilde{\Delta}_S = (\mathbf{\Gamma}_{SS})^{-1}(-\mathbf{R}_{1,SS}\theta^*_S - \lambda\tilde{z}).$$

Note that by assumption, $\mathbf{\Gamma}_{SS}$ is invertible. We obtain that

$$\begin{aligned}
\|\tilde{\Delta}_S\|_\infty &\leq |||(\mathbf{\Gamma}_{SS})^{-1}|||_\infty \Big[\|\mathbf{R}_{1,SS}\theta^*_S\|_\infty + \|r_2\|_\infty + \lambda\Big] \\
&< |||(\mathbf{\Gamma}_{SS})^{-1}|||_\infty \Big[|||\mathbf{\Theta}^*_{\text{wide}}|||_\infty \|\text{vec}(\mathbf{R}_1)\|_\infty + \|r_2\|_\infty + \lambda\Big] \\
&\leq |||(\mathbf{\Gamma}_{SS})^{-1}|||_\infty \times \frac{(6-\alpha)}{3(2-\alpha)}\lambda. \qquad (7.7)
\end{aligned}$$

Since $\|\mathbf{R}_1\|_\infty < \epsilon_1$, we have $|||\mathbf{R}_{1,SS}|||_\infty \leq d\epsilon_1 < 1/c_{\mathbf{\Gamma}^*}$. This implies that

$$|||(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{R}_{1,SS}|||_\infty \leq |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty |||\mathbf{R}_{1,SS}|||_\infty < 1,$$

which gives us the following bound in the error in the inverse in the matrix ℓ_∞ norm,

$$\begin{aligned}
|||(\mathbf{\Gamma}_{SS})^{-1} - (\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty &\leq \frac{|||(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{R}_{1,SS}|||_\infty}{1 - |||(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{R}_{1,SS}|||_\infty} \times |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty \\
&\leq \frac{|||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty |||\mathbf{R}_{1,SS}|||_\infty}{1 - |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty |||\mathbf{R}_{1,SS}|||_\infty} \times |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty.
\end{aligned}$$

Application of the triangle inequality, along with our definition of

$$c_{\mathbf{\Gamma}^*} = |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty,$$

yields

$$\begin{aligned}
\left|\left|\left|(\boldsymbol{\Gamma}_{SS})^{-1}\right|\right|\right|_\infty &\leq \left|\left|\left|(\boldsymbol{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty + \left|\left|\left|(\boldsymbol{\Gamma}_{SS})^{-1} - (\boldsymbol{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \\
&= \left|\left|\left|(\boldsymbol{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \times \frac{1}{1 - \left|\left|\left|(\boldsymbol{\Gamma}^*_{SS})^{-1}\right|\right|\right|_\infty \left|\left|\left|\mathbf{R}_{1,SS}\right|\right|\right|_\infty} \\
&\leq \frac{c_{\boldsymbol{\Gamma}^*}}{1 - d c_{\boldsymbol{\Gamma}^*} \epsilon_1} \\
&\leq \frac{c_{\boldsymbol{\Gamma}^*}}{1 - \alpha/6}, \qquad (7.8)
\end{aligned}$$

where the last inequality uses the assumption that $d\epsilon_1 \leq \alpha/6c_{\boldsymbol{\Gamma}^*}$. Substituting (7.8) into (7.7), it is straightforward to show that $G_2 < \alpha/3$. Therefore, $G_1 + G_2 + G_3 < \alpha$, which yields that $\|\tilde{z}_{S^c}\| < 1$.

Along the way we have also proven the second part of the claim. Indeed, from (7.7) and (7.8), we have

$$\|\tilde{\Delta}_S\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}^*}}{1 - \alpha/6} \times \frac{(6 - \alpha)}{3(2 - \alpha)} \lambda = \frac{2 c_{\boldsymbol{\Gamma}^*} \lambda}{2 - \alpha}.$$

7.2. Proof of Corollary 1

We need to show that the conditions in Theorem 1, specifically those in (6.6), hold with the claimed probability. Since $r_2 = g - g^* = \mathrm{vec}(\mathbf{I}_{m\times m}) - \mathrm{vec}(\mathbf{I}_{m\times m}) = 0$, the second inequality in (6.6) can be trivially satisfied with any $\epsilon_2 > 0$. Thus, we only need to show that we can bound $\|\mathbf{R}_1\|_\infty$ by some suitable ϵ_1 with sufficiently large probability. To do so, we apply a Bernstein-type concentration inequality for the entries of W that is also used by Ravikumar et al. (2011). Lemma B.1 below states the inequality, as given in their paper.

The matrix $\mathbf{R}_1$ features only entries in $\mathbf{W} - \boldsymbol{\Sigma}^*$. By taking a union bound over the m^2 entries of $\mathbf{W}$, plugging in our lower bound for n and observing that $\sigma = 1$ in the Gaussian case, Lemma B.1 yields that

$$\Pr\left[\|\mathbf{R}_1\|_\infty \geq \sqrt{\frac{c^*(\log m^{\tau_1} + \log 4)}{n}}\right] \leq \exp\{-\log m^{\tau_1} + 2\log m\} = \frac{1}{m^{\tau_1 - 2}}.$$

In addition, each row in $\left|\left|\left|\mathbf{R}_{\cdot S}\right|\right|\right|_\infty$ features at most d entries from the matrix $\mathbf{W} - \boldsymbol{\Sigma}^*$. Hence, it follows from another union bound, and choosing n at least

$$c^* c_1^2 d^2 (\log m^{\tau_1} + \log 4)$$

where c^* and c_1 are defined in the corollary statement, that

$$\Pr\left[\left|\left|\left|\mathbf{R}_{\cdot S}\right|\right|\right|_\infty > \frac{1}{c_1}\right] \leq \frac{1}{m^{\tau_1 - 2}}.$$

Thus, applying Theorem 1 with

$$\epsilon_1 = \sqrt{\frac{c^*(\log m^{\tau_1} + \log 4)}{n}}$$

shows that our choices for λ and n give the high probability statement in Corollary 1.

When looking back at the proof of Theorem 1, we see that as a consequence of having $r_2 = 0$, we need only be concerned with bounding terms G_1 and G_2. We may thus bound G_1 and G_2 each by $\alpha/2$ instead of $\alpha/3$ and ignore the G_3 term entirely, as it is 0. This leads us to having $c_1 = (4/\alpha)c_{\mathbf{\Gamma}^*}$, as opposed to the expected $(6/\alpha)c_{\mathbf{\Gamma}^*}$.

7.3. *Proof of Corollary 2*

We proceed as for the proof of Corollary 1 and use concentration results to satisfy the bounds from (6.6) in Theorem 1. However, we now bound $\|\mathbf{R}_1\|_\infty$ and $\|r_2\|_\infty$ using concentration inequalities for general log-concave measures (any truncated multivariate normal density is log-concave).

Let $X^{(i)} = (X_{i1}, \ldots, X_{im})$ be i.i.d. according to $N(0, (\mathbf{K}^*)^{-1})$ with truncation to $\mathbb{R}_+^m$. Take

$$\epsilon_1 = \frac{\left[\left(\frac{L}{2}\right)(\log m^{\tau_2} + \log 2)\right]^4}{\sqrt{n}} \sqrt{\max_j \operatorname{Var}[X_j^4]}, \tag{7.9}$$

$$\epsilon_2 = \frac{\left[\left(\frac{L}{2}\right)(\log m^{\tau_2} + \log 2)\right]^2}{\sqrt{n}} \sqrt{\max_j \operatorname{Var}[X_j^2]}. \tag{7.10}$$

From Lemma B.3 below, we know that for the absolute constant L specified in Lemma B.2, we have,

$$\begin{aligned}
&\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_{ij}X_{ik}X_{i\ell}^2 - E[X_jX_kX_\ell^2]\right| > \epsilon_1\right] \\
&\quad < \exp\left\{-\frac{2}{L}\left(\frac{\sqrt{n}\epsilon_1}{\sqrt{\max_{j,k,\ell} \operatorname{Var}[X_jX_kX_\ell^2]}}\right)^{\frac{1}{4}}\right\}, \\
&\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_{ij}X_{ik} - E[X_jX_k]\right| > \epsilon_2\right] \\
&\quad < \exp\left\{-\frac{2}{L}\left(\frac{\sqrt{n}\epsilon_2}{\sqrt{\max_{j,k,\ell} \operatorname{Var}[X_jX_k]}}\right)^{\frac{1}{2}}\right\}
\end{aligned}$$

for all $j, k, \ell = 1, \ldots, m$. By a union bound over no more than $2m^3$ events, we have both $\|\mathbf{R}_1\|_\infty < \epsilon_1$ and $\|r_2\|_\infty < \epsilon_2$ with probability at least $1 - 1/m^{\tau_2 - 3}$ as $m \to \infty$. Applying Theorem 1 with the chosen ϵ_1 and ϵ_2 thus shows that our choices for λ and n lead to the claim in Corollary 2.

8. Discussion

This paper proposes the use of regularized score matching for estimation of conditional independence graphs in high dimensions. The focus is on modifying the score matching loss of Hyvärinen (2005) with an ℓ_1 penalty to accommodate underlying sparsity, which is in the spirit of popular existing methods such as glasso and neighborhood selection. This said, any other regularization scheme can be considered instead. For instance, the method from Defazio and Caetano (2012) can be applied to encourage hub structure in the inferred graph.

Our study of the Gaussian example of Meinshausen (2008) suggests that ℓ_1-regularized score matching falls in between neighborhood selection and glasso in terms of conditions for required for graph selection consistency. Here, the glasso requires the most stringent conditions, and the score matching approach appears to be similar to pseudo-likelihood methods that work with symmetric estimates of precision matrices, such as SPACE (Peng et al., 2009) and subsequent reformulations such as CONCORD (Khare, Oh and Rajaratnam, 2015). However, regularized score matching is particularly convenient in that the score matching loss is a quadratic function, even for non-Gaussian exponential families. This brings about piecewise linear solution paths and allows for a simple theoretical analysis. We anticipate that the simple structure of score matching will lead to further advances in graphical modeling, such as computationally efficient techniques to deal with corrupted or missing data, in the spirit of Loh and Wainwright (2012), or new methods to tune regularization parameters, as in Chichignoud, Lederer and Wainwright (2014).

Regularized score matching is an interesting method for Gaussian models, as we showed empirically and theoretically. In particular, for consistency (under the usual irrepresentability conditions), the sample n must be on the order $\Omega(d^2 \log m)$, which matches the conditions for the existing methods mentioned above. However, as our simulation study shows, regularized score matching really shines in the context of non-Gaussian models, where it eliminates the need to deal with computationally intractable normalization constants in a way that the loss continues to be a quadratic function of parameters. This opens a lot of new possibilities for graphical modeling such as the truncated normal model we applied to RNAseq data.

Score matching applies to continuous data. While Hyvärinen (2007) discusses a ratio matching method for discrete data, it is not as computationally convenient as its continuous counterpart. A different approach of adding Gaussian noise to discrete data was proposed for imaging problems by Kingma and LeCun (2010). Exploring the merits of their approach for graphical modeling, and supplying supporting theory, would be an interesting problem for future work.

Appendix A: Implementation

The piecewise linear solution path for regularized score matching can be computed using Algorithm 1, which is an adaptation of the LARS-Lasso algorithm

Algorithm 1

Initialize $\theta = 0$
Initialize $\hat{S} = \arg\max_j \left|(\mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x}))_j\right|$
Initialize $\xi_{\hat{S}} = -\text{sign}\left((\mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x}))_{\hat{S}}\right)$
Initialize $\xi_{\hat{S}^c} = 0$
while $\|\mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x})\|_\infty > 0$ and $\mathbf{\Gamma}_{\hat{S}\hat{S}}$ is invertible **do**
  $\eta_1 \leftarrow \min\{\eta > 0 : |(\mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x})|_j = |(\mathbf{\Gamma}(\mathbf{x})\theta + g(\mathbf{x})|_{\hat{S}}, j \notin \hat{S}\}$.
  $\eta_2 \leftarrow \min\{\eta > 0 : (\theta + \eta\boldsymbol{\xi})_j = 0, j \in \hat{S}\}$.
  $\eta \leftarrow \min\{\eta_1, \eta_2\}$.
  $\theta \leftarrow \theta + \eta\xi$
  if $\eta = \eta_1$ **then**
    Add variable that attains equality to $\hat{S}$.
  else
    Remove variable that attains 0 from $\hat{S}$.
  end if
  $\xi_{\hat{S}} \leftarrow \left(\mathbf{\Gamma}(\mathbf{x})_{\hat{S}\hat{S}}\right)^{-1} \text{sign}(\theta_{\hat{S}})$
end while

for linear regression (Efron et al., 2004). It is also a special case of the algorithm found in Rosset and Zhu (2007). In our pseudocode, $\hat{S}$ is the current active set, i.e., $\hat{S} = \{j : \theta_j^\lambda \neq 0\}$ for the currently relevant value of the regularization parameter λ.

In the Gaussian and truncated Gaussian case, the algorithm stops when the active set has size $|\hat{S}| = \min\{n, m\}m$. For larger active sets the matrix $\mathbf{\Gamma}_{\hat{S}\hat{S}}$ is not invertible. Finding the step size in Algorithm 1 requires $\mathcal{O}\left(\min\{n, m\}m\right)$ operations, while the inversion step is at its worst $\mathcal{O}(|\hat{S}|^2) = \mathcal{O}\left(\min\{n, m\}^2 m^2\right)$. Overall, the complexity of Algorithm 1 can be found to be $\mathcal{O}\left(\min\{n, m\}^3 m^2\right)$; the heaviest cost comes from the matrix inversion step.

For large-scale problems, LARS-type algorithms may be slow and coordinate-descent methods are popular alternatives (see e.g. Friedman et al., 2007). Algorithm 2 describes a coordinate-descent algorithm to minimize the regularized score matching objective from (3.1). It entails updating one coordinate, or one element in the parameter vector/matrix, such that it minimizes the objective function while holding all others as constant, until a convergence criterion is satisfied. Results in Tseng (2001) ensure convergence of Algorithm 2.

Example 1 (cont.). For the Gaussian case, the coordinate descent procedure alternates between updating the diagonal entries and off-diagonal entries, by manipulating the estimating equations (3.10) and (3.11) accordingly. The updates are of the form

$$\kappa_{jj}^{(t+1)} \leftarrow \frac{1 - \sum_{j' \neq j} w_{jj'} \kappa_{jj'}^{(t)}}{w_{jj}},$$

$$\kappa_{jk}^{(t+1)}, \kappa_{kj}^{(t+1)} \leftarrow \text{Soft}\left(\frac{-\sum_{j' \neq j} w_{jj'} \kappa_{j'k}^{(t)} - \sum_{k' \neq k} w_{jk'} \kappa_{k'k}^{(t)}}{w_{jj} + w_{kk}}, \frac{2\lambda}{w_{jj} + w_{kk}}\right),$$

for $j, k \in \{1, \ldots, m\}$. The computational complexity of this scheme can be shown to be $\min(\mathcal{O}(nm^2), \mathcal{O}(m^3))$, which is the same as for the methods classified

Algorithm 2

Input: Initial estimate $\hat{\theta}^{(0)}$
Input: t_{max}, maximum number of iterations
Input: ϵ, the maximal tolerance level
Initialize $t \leftarrow 1$
Initialize $C \leftarrow \epsilon + 1$ (C stands for convergence criteria)
while $C > \epsilon$ or $t < t_{max}$ **do**
  $\hat{\theta}^{(t)} \leftarrow \hat{\theta}^{(t-1)}$
  for $j \leftarrow 1, 2, \ldots, s$ **do**
    $\hat{\theta}_j^{(t)} \leftarrow \text{Soft}\left(\frac{-\left(\mathbf{\Gamma}(\mathbf{x})_{-j,j}\right)^T \hat{\theta}_{-j}^{(t)} - g(\mathbf{x})_j}{\mathbf{\Gamma}(\mathbf{x})_{jj}}, \frac{\lambda}{\mathbf{\Gamma}(\mathbf{x})_{jj}}\right)$.
  end for
  $C \leftarrow \|\hat{\theta}^{(t)} - \hat{\theta}^{(t-1)}\|_1$
  $t \leftarrow t + 1$
end while

under SPACE; the complexity of glasso is $\mathcal{O}(m^3)$. We do not prove this fact, as it follows directly from reasoning elaborated on in Khare, Oh and Rajaratnam (2015).

Appendix B: Concentration results

Corollaries 1 and 2 make use of the following concentration results. The first lemma is used to prove Corollary 1 while the latter two (one is derived from the other) are used to prove Corollary 2.

Lemma B.1 (Ravikumar et al., 2011). *If* $\left(X_1, \ldots, X_m\right)$ *is a zero-mean random vector with covariance matrix* $\mathbf{\Sigma}^*$ *such that* $X_i/\sqrt{\Sigma_{ii}^*}$ *is sub-Gaussian with scale parameter* σ*, then the sample covariance matrix* $\mathbf{W}$*, for* n *i.i.d. samples, satisfies the bound*

$$\Pr[|\mathbf{W}_{jk} - \mathbf{\Sigma}_{jk}^*| > \delta] \leq 4 \exp\left\{-\frac{n\delta^2}{128(1+4\sigma^2)^2 \max\limits_{j=1,\ldots,m}(\mathbf{\Sigma}_{jj}^*)^2}\right\} \tag{B.1}$$

for any fixed choice of two indices $1 \leq j, k \leq m$ *and for all* $\delta \in (0, 40 \max\limits_{j=1,\ldots,m} \mathbf{\Sigma}_{jj}^*)$.

Lemma B.2 (Carbery and Wright, 2001). *Let* $\mathcal{X}$ *be a Banach space, and let* $f : \mathbb{R}^m \to \mathcal{X}$ *be a polynomial of degree at most* z*. Suppose* $0 < \zeta_1 \leq \zeta_2 < \infty$ *and* μ *is a log-concave probability measure on* $\mathbb{R}^m$*. Then*

$$\left(\int \|f(x)\|^{\zeta_2/z} d\mu(x)\right)^{1/\zeta_2} \leq L \frac{\max(\zeta_2, 1)}{\max(\zeta_1, 1)} \left(\int \|f(x)\|^{\zeta_1/z} d\mu(x)\right)^{1/\zeta_1}, \tag{B.2}$$

where $L > 0$ *is an absolute constant.*

From this lemma we may derive the following concentration result. After proving the lemma, we comment on how it is used in the proof of Corollary 2.

Lemma B.3. *Consider a degree z polynomial $f(X) = f(X_1, \ldots, X_m)$, where $X_1, \ldots, X_m$ are possibly dependent random variables with log-concave joint distribution on $\mathbb{R}^m$. Let $L > 0$ be the constant from Lemma B.2. Then, for all δ such that*

$$K := \frac{2}{L}\left(\frac{\delta}{e\sqrt{\mathrm{Var}[f(X)]}}\right)^{1/z} \geq 2, \tag{B.3}$$

we have,

$$\Pr[|f(X) - E[f(X)]| > \delta] \;\leq\; \exp\left\{-\frac{2}{L}\left(\frac{\delta}{\sqrt{\mathrm{Var}[f(X)]}}\right)^{1/z}\right\}. \tag{B.4}$$

Proof. Choosing $\zeta_1 = 2z$ and $\zeta_2 = Kz$ in Lemma B.2, we have

$$E[|f(X) - E[f(X)]|^K]^{\frac{1}{K}} \;\leq\; \left(\frac{LK}{2}\right)^z \sqrt{\mathrm{Var}[f(X)]}.$$

Hence, by Markov's inequality, for any δ satisfying (B.3),

$$\begin{aligned} P[|f(X) - E[f(X)]| > \delta] \;&\leq\; \frac{E[|f(X) - E[f(X)]|^K]}{\delta^K} && \text{(B.5)}\\ &\leq\; \left[\left(\frac{LK}{2}\right)^z \frac{\sqrt{\mathrm{Var}[f(X)]}}{\delta}\right]^K && \text{(B.6)}\\ &=\; \exp\{-K\} && \text{(B.7)}\\ &=\; \exp\left\{-\frac{2}{L}\left(\frac{\delta}{\sqrt{\mathrm{Var}[f(X)]}}\right)^{\frac{1}{z}}\right\}, && \text{(B.8)} \end{aligned}$$

and the proof is complete. □

In the proof of Corollary 2, we apply Lemma B.3 with $\delta = \epsilon_1$ from (7.9) and with $\delta = \epsilon_2$ from (7.10). It thus needs to be checked that condition (B.3) holds in these two cases. Indeed, the condition holds as long as

$$m \;\geq\; \exp\left\{\frac{2\sqrt{e} - \log 2}{\tau_2}\right\}. \tag{B.9}$$

To see this, we substitute ϵ_1 and ϵ_2 for δ in (B.3), take $z = 4$ and 2 respectively, to find a term that is lower bounded by $(\tau_2 \log m + \log 2)/e^2$. Here, the $1/\sqrt{n}$ factor in ϵ_1 and ϵ_2 cancels out with the $1/\sqrt{n}$ term generated by the $\sqrt{\mathrm{Var}[f(X)]}$ term in the denominator. (Recall that in our scenario $f(X)$ is an empirical average). The more stringent condition on m comes from ϵ_2 and is stated in (B.9). Thus, if (B.9) holds, (B.3) is satisfied. Since $\tau_2 > 3$, the right-hand side of (B.9) never exceeds

$$\exp\left\{\frac{1}{3}(2\sqrt{e} - \log 2)\right\} < 3.$$

Hence, in our application of Lemma B.3, the condition from (B.3) holds for $m \geq 3$.

Appendix C: Experiments

We perform experiments, similar to those found in related work, that give empirical support for Corollary 1. This corollary treats Gaussian graphical models for which the sample size n ought to be of order $d^2 \log m$. We experiment by varying the number of variables m, the degree d, and the minimum signal strength. Following Ravikumar et al. (2011), we define the 'model complexity' to be

$$C := \frac{4}{\alpha} c_{\mathbf{\Gamma}^*} \times \max_j \mathbf{\Sigma}^*_{jj}. \tag{C.1}$$

In addition, we investigate how the sample size n required for sparsistency for non-negative Gaussian graphical models needs to depend on m. All reported results are based on averaging over 100 trials.

C.1. Gaussian experiments

We conduct our experiments using three graph structures: (a) a chain, (b) a 2-D lattice with 4 nearest neighbors, and (c) a star. We consider (a) and (b) when varying the number of variables m, in which case we vary the length of the chain and the number of nodes in the lattice. This keeps the degree d constant. The effect that d has on the sample complexity is investigated using stars. We let the regularization parameter λ scale with $\sqrt{\log m/n}$, a choice corroborated by Corollary 1.

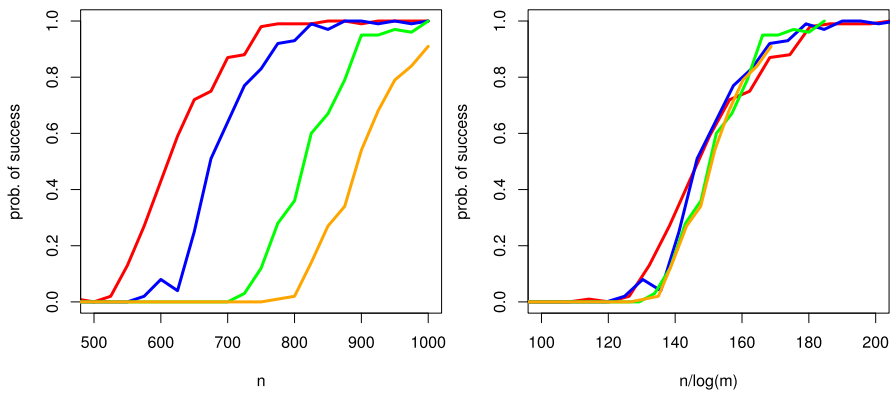
Dependence on number of nodes

Consider first the case where the underlying conditional independence graph is a chain of length $m \in \{64, 100, 225, 375\}$. The degree d is always 2, and we choose the tridiagonal precision matrix $\mathbf{K}^*$ to have entries $\kappa^*_{jk} = 0.3$ if $(j,k) \in E$ and $\kappa^*_{jj} = 1$ for $j = 1, \ldots m$. Here, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$ are constant across all m.

Figure 10 shows the probability of correct signed support recovery plotted against the sample size n, with different curves corresponding to different m. As expected, we see from Figure 10(a) that successful support recovery requires n to grow with m. However, upon rescaling n by $1/\log m$, the curves overlap as seen in Figure 10(b).

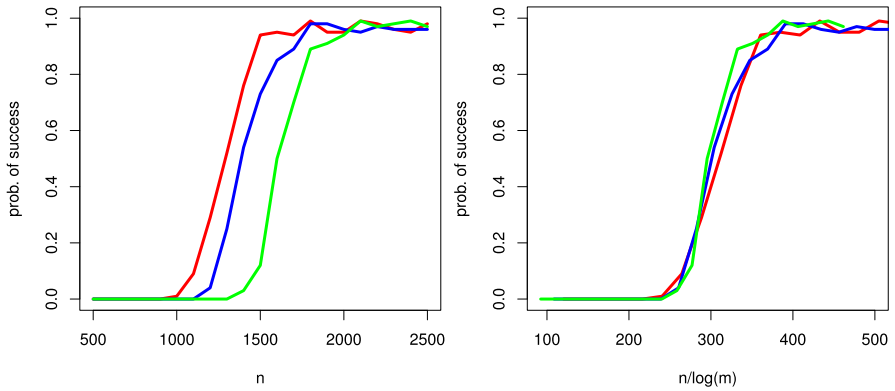
We repeat the experiment with the 2-D lattice graph with $m \in \{64, 100, 225\}$ nodes. Each node is connected to four nearest neighbors such that the degree d is always 4. We choose $\mathbf{K}^*$ with $\kappa^*_{jk} = 0.2$ for $(j,k) \in E$ and $\kappa^*_{jj} = 1$ for $j = 1, \ldots m$. Again, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$ are constant across all m. The results are presented in Figure 11, which shows curves of recovery probabilities that stack on top of one another when n by $1/\log m$.

We conclude that with C and d held constant, the sample size n needs to scale with $\log m$ for consistent signed support recovery. This is consistent with Corollary 1.



(a) (b)

FIG 10. *Relative frequencies of signed support recovery for Gaussian observations with a conditional independence graph that is a chain of varying length m. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to* $m = 64$ *(——),* $m = 100$ *(——),* $m = 225$ *(——) and* $m = 375$ *(——).*

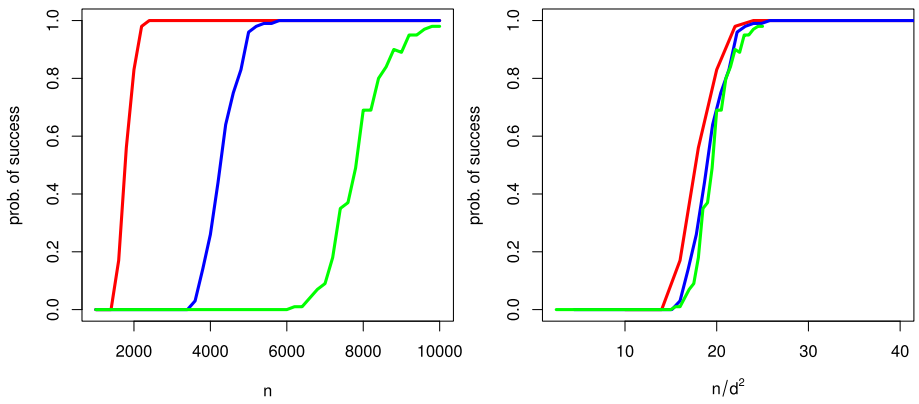


(a) (b)

FIG 11. *Relative frequencies of signed support recovery for Gaussian observations whose conditional independence graph is a 4-nearest neighbor lattice with m nodes. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to* $m = 64$ *(——),* $m = 100$ *(——), and* $m = 225$ *(——).*

Dependence on node degree

We now fix the number of nodes to $m = 200$ and vary d. We consider a star graphs with varying hub node degree $d \in \{15, 20, 25\}$. The precision matrix $\mathbf{K}^*$ is chosen such that $\sigma^*_{jk} = 2.5/d$ for $(j,k) \in E$, and $\sigma^*_{jj} = 1$ for $j = 1, \ldots m$. Now, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$ are constant across all d.



(a) (b)

FIG 12. *Relative frequencies of signed support recovery for Gaussian observations whose conditional independence graph is a star with varying degree d. Panels (a) and (b) differ only in the scaling of the x-axis. The colored lines correspond to* $d = 10$ *(——),* $d = 15$ *(——), and* $d = 20$ *(——).*

Figure 12 shows the probability of correct signed support recovery plotted against n. The left panel demonstrates that correct recovery is more difficult with increasing d. Larger n is needed to attain the same success rate. Upon rescaling n by $1/d^2$ in the right panel, the three curves align. This validates Corollary 1 in that for fixed m, α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$, the sample size n needs to scale with d^2 to ensure sign consistency.

Dependence on 'model complexity'

We return to the chain-structured graphs considered earlier in this section. This time, however, we fix $m = 64$ and $d = 2$ while changing the edge strengths κ^*_{jk} for $(j, k) \in E$, which alters C from (C.1). We plot the probability of correct signed support recovery against n for varying C. In the resulting Figure 13, the curves shift right as C becomes larger so a larger n is needed to attain the same probability of correct signed support recovery when C grows. This is again consistent with the implications of Corollary 1. We do not believe that the lower bound we found for n is sharp enough in terms of its dependence on α, $c_{\mathbf{K}^*}$ and $c_{\mathbf{\Gamma}^*}$ to determine the rescaling we must perform on n to align the curves.

C.2. Non-negative Gaussian experiments

Finally, we experiment with regularized non-negative score matching for normal observations truncated to the positive orthant. According to Corollary 2, a sample size of $n = \Omega(d^2(\log m)^8)$ is sufficient for signed support recovery. The

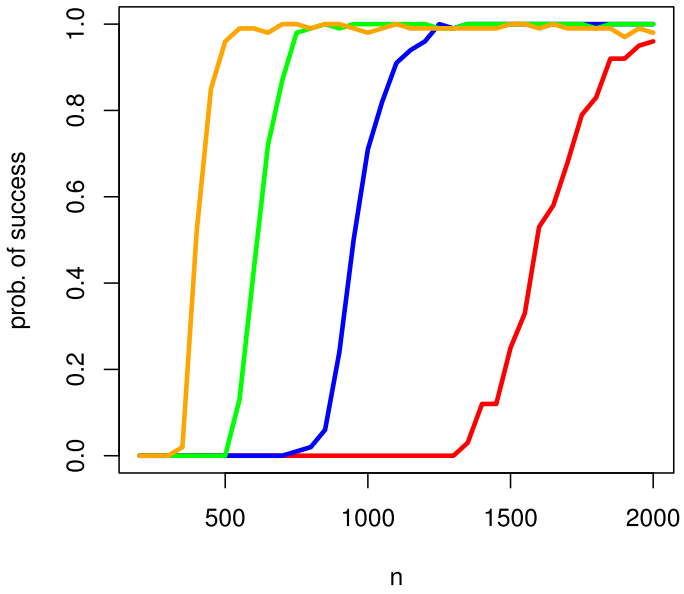


FIG 13. *Relative frequencies of signed support recovery for Gaussian observations whose conditional independence graph is a chain of fixed length. The different curves correspond to different signal strength summarized in the model complexity C. The colored lines correspond to $C = 857$ (——), $C = 668$ (——), $C = 576$ (——) and $C = 543$ (——).*

aim of our experiments is to explore to what extent this scaling is necessary. Specifically, we will consider exponents other than 8 for $\log m$.

For our experiments, we revisit the chain-structured graphs from Section C.1 and choose a triangular matrix $\mathbf{K}^*$ with $\kappa^*_{jk} = 0.3$ if $(j,k) \in E$ and and $\kappa^*_{jj} = 1$ for $j = 1, \ldots m$. The degree d is fixed at 2 and we only vary $m \in \{20, 25, 30\}$. We let the regularization parameter λ to scale with $\sqrt{(\log m)^8/n}$. Figure 14 plots the probability of correct signed support recovery against n, with different curves for the different values of m.

Panel (a) in Figure 14 illustrates that, larger n is needed account for larger m. The other three panels have the x-axis rescaled to $n/(\log m)^a$ for exponents $a \in \{6, 7, 8\}$. Panel (b) suggests that n scaling with $(\log m)^6$ is not sufficient for support recovery. Comparing panels (c) and (d), $(\log m)^8$ seems more than what is necessary. It thus appears that the scaling of the sample size we assumed in Corollary 2 is suboptimal but not drastically so.

Supplementary Material

Computer code

(doi: 10.1214/16-EJS1126SUPP; .zip). This supplemental material provides R code that we used for our numerical experiments.

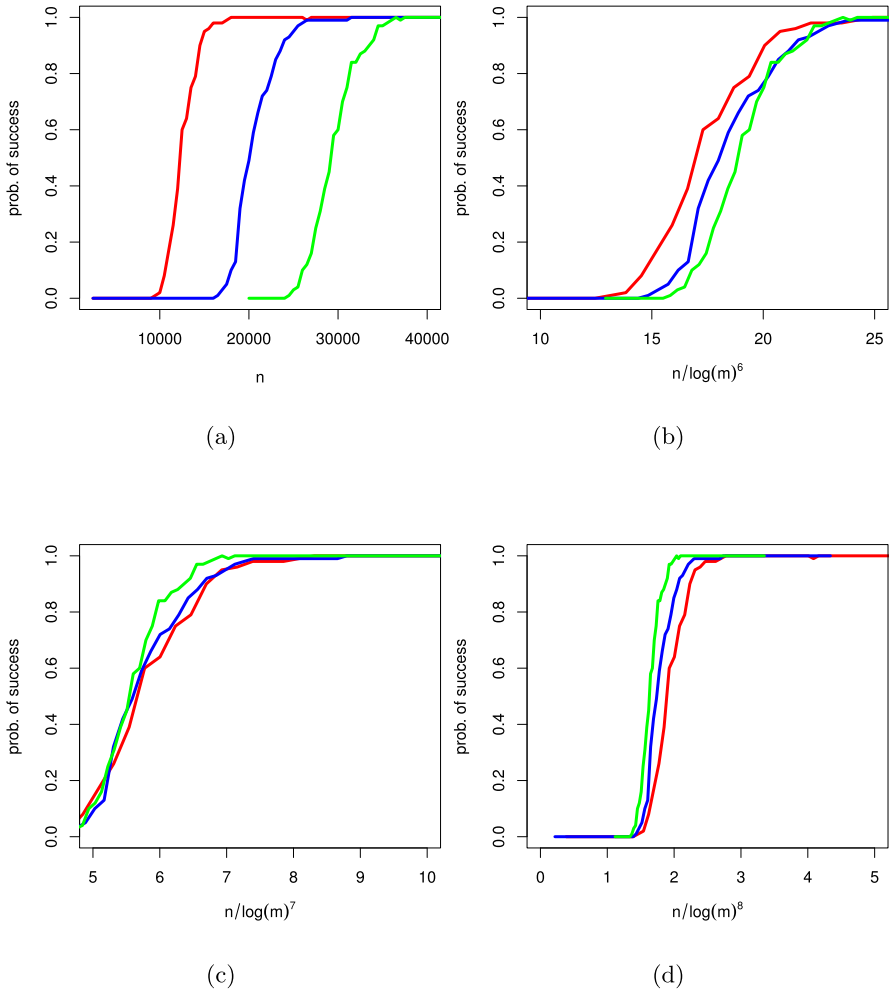


FIG 14. *Relative frequencies of signed support recovery for truncated Gaussian observations whose conditional independence graph is a chain of varying length m. The four panels differ only in the scaling of the x-axis. The colored lines correspond to* $m = 20$ *(——), m* $= 25$ *(——), and m* $= 30$ *(——).*

References

AHONEN, T. J., XIE, J., LEBARON, M. J., ZHU, J., NURMI, M., ALANEN, K., RUI, H. and NEVALAINEN, M. T. (2003). Inhibition of transcription factor Stat5 induces cell death of human prostate cancer cells. *Journal of Biological Chemistry* **278** 27287–27292.

ALBERT, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science* **118** 4947–4957.

Allen, G. I. and Liu, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. NanoBioscience* **12** 189–198.

Arnold, B. C., Castillo, E. and Sarabia, J. M. (1999). *Conditional specification of statistical models*. Springer-Verlag, New York. MR1716531

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

Barber, R. F. and Drton, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.* **9** 567–607. MR3326135

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer, Heidelberg. MR2807761 (2012e:62006)

Carbery, A. and Wright, J. (2001). Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Math. Res. Lett.* **8** 233–248. MR1839474 (2002h:26033)

Carter, S. L., Brechbühler, C. M., Griffin, M. and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20** 2242–2250.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika* **95** 759–771. MR2443189

Chichignoud, M., Lederer, J. and Wainwright, M. (2014). Tuning Lasso for sup-norm optimality. arXiv:1410.0247. MR3217454

Dawid, A. P. and Musio, M. (2013). Estimation of spatial processes using local scoring rules. *AStA Adv. Stat. Anal* **97** 173–179. MR3045766

Defazio, A. and Caetano, T. S. (2012). A convex formulation for learning scale-free networks via submodular relaxation. *Adv. Neural Inf. Process. Syst.* 1250–1258.

Dempster, A. P. (1972). Covariance selection. *Biometrics* 157–175.

Dobra, A. and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statist. Sci.* **22** 430–449. MR2416818

Edwards, D. (2000). *Introduction to graphical modelling*, Second ed. Springer-Verlag, New York. MR1880319

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166 (2005d:62116)

Fan, S., Meng, Q., Auborn, K., Carter, T. and Rosen, E. (2006). BRCA1 and BRCA2 as molecular targets for phytochemicals indole-3-carbinol and genistein in breast and prostate cancer cells. *Brit. J. Cancer* **94** 407–426.

Fellinghauer, B., Bühlmann, P., Ryffel, M., von Rhein, M. and Reinhardt, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comput. Statist. Data Anal.* **64** 132–152. MR3061894

Finegold, M. and Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.* **5** 1057–1080. MR2840186 (2012i:62151)

Forbes, P. G. M. and Lauritzen, S. (2015). Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra Appl.* **473** 261–283. MR3338335

Foygel, R. and Drton, M. (2010a). Exact block-wise optimization in group lasso for linear regression. arXiv:1010.3320.

Foygel, R. and Drton, M. (2010b). Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Inf. Process. Syst.* **23** 2020–2028.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models Technical Report, Stanford University.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

Gao, X., Pu, D. Q., Wu, Y. and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statist. Sinica* **22** 1123–1146. MR2987486

Gayther, S. A., de Foy, K. A., Harrington, P., Pharoah, P., Dunsmuir, W. D., Edwards, S. M., Gillett, C., Ardern-Jones, A., Dearnaley, D. P., Easton, D. F. et al. (2000). The frequency of germ-line mutations in the breast cancer predisposition genes BRCA1 and BRCA2 in familial prostate cancer. *Cancer Res.* **60** 4513–4518.

Gelman, A. and Meng, X.-L. (1991). A note on bivariate distributions that are conditionally normal. *Amer. Statist.* **45** 125–126.

Gu, L., Vogiatzi, P., Puhr, M., Dagvadorj, A., Lutz, J., Ryder, A., Addya, S., Fortina, P., Cooper, C., Leiby, B. et al. (2010). Stat5 promotes metastatic behavior of human prostate cancer cells in vitro and in vivo. *Endocr. Relat. Cancer* **17** 481–493.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. et al. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430** 88–93.

Höfling, H. and Tibshirani, R. J. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** 695–709. MR2249836

Hyvärinen, A. (2007). Some extensions of score matching. *Comput. Statist. Data Anal.* **51** 2499–2512. MR2338984

Jalali, A., Ravikumar, P. D., Vasuki, V. and Sanghavi, S. (2011). On learning discrete graphical models using group-sparse regularization. In *AISTATS 2011* 378–387.

Jeong, H., Mason, S. P., Barabási, A.-L. and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411** 41–42.

Khare, K., Oh, S.-Y. and Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. Roy. Statist. Soc. Ser. B* **77** 803–825. MR3382598

Kingma, D. P. and LeCun, Y. (2010). Regularized estimation of image statis-

tics by score matching. In *Adv. Neural Inf. Process. Syst.* 1126–1134.

Kishi, H., Igawa, M., Kikuno, N., Yoshino, T., Urakami, S. and Shiina, H. (2004). Expression of the survivin gene in prostate cancer: correlation with clinicopathological characteristics, proliferative activity and apoptosis. *J. Urology* **171** 1855–1860.

Köster, U. and Hyvärinen, A. (2007). A two-layer ICA-like model estimated by score matching. In *ICANN 2007* 798–807. Springer.

Lauritzen, S. L. (1996). *Graphical models* **17**. Oxford University Press. MR1419991

Le, Q. V., Karpenko, A., Ngiam, J. and Ng, A. Y. (2011). ICA with reconstruction cost for efficient overcomplete feature learning. In *Adv. Neural Inf. Process. Syst.* 1017–1025.

Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Mol. Syst. Biol.* **4** 213.

Lee, S.-I., Ganapathi, V. and Koller, D. (2007). Efficient structure learning of Markov networks using ℓ_1-regularization. In *Advances in Neural Information Processing Systems 19* (B. Schölkopf, J. C. Platt and T. Hoffman, eds.) 817–824. MIT Press.

Lin, L., Drton, M. and Shojaie, A. (2016). Supplement to "Estimation of high-dimensional graphical models using regularized score matching".

Liu, H., Han, F. and Zhang, C.-h. (2012). Transelliptical graphical models. In *Adv. Neural Inf. Process. Syst.* 809–817.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Adv. Neural Inf. Process. Syst.* 1432–1440.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

Meinshausen, N. (2008). A note on the Lasso for Gaussian graphical model selection. *Statist. Probab. Lett.* **78** 880–884. MR2398362

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 (2008b:62044)

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. Roy. Statist. Soc. Ser. B* **72** 417–473. MR2758523

Mitra, A., Fisher, C., Foster, C., Jameson, C., Barbachanno, Y., Bartlett, J., Bancroft, E., Doherty, R., Kote-Jarai, Z., Peock, S. et al. (2008). Prostate cancer in male BRCA1 and BRCA2 mutation carriers has a more aggressive phenotype. *Brit. J. Cancer* **98** 502–507.

Miyamura, M. and Kano, Y. (2006). Robust Gaussian graphical modeling. *J. Multivariate Anal.* **97** 1525–1550. MR2275418

Moser, C., Ruemmele, P., Gehmert, S., Schenk, H., Kreutz, M. P., Mycielska, M. E., Hackl, C., Kroemer, A., Schnitzbauer, A. A., Stoeltzing, O. et al. (2012). STAT5b as molecular target in pancreatic cancer—Inhibition of tumor growth, angiogenesis, and metastases. *Neoplasia* **14** 915–IN12.

Okamoto, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.* **1** 763–765. MR0331643 (48 ##9975)

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 (2011d:62066)

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

Rocha, G. V., Zhao, P. and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE) Technical Report, University of California, Berkeley.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012–1030. MR2341696 (2009b:62140)

Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *ICML* 848–855.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *J. Roy. Statist. Soc. Ser. B* **75** 55–80. MR3008271

Shariat, S. F., Lotan, Y., Saboorian, H., Khoddami, S. M., Roehrborn, C. G., Slawin, K. M. and Ashfaq, R. (2004). Survivin expression is associated with features of biologically aggressive prostate carcinoma. *Cancer* **100** 751–757.

Shojaie, A. and Sedaghat, N. (2016). How similar are estimated networks of different cancer subtypes? In *Big and Complex Data Analysis: Statistical Methodologies and Applications* (S. E. Ahmed, ed.) Springer, New York.

Sun, H. and Li, H. (2012). Robust Gaussian graphical modeling via ℓ_1 penalization. *Biometrics* **68** 1197–1206. MR3040026

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375

Tryggvadóttir, L., Vidarsdóttir, L., Thorgeirsson, T., Jonasson, J. G., Ólafsdóttir, E. J., Ólafsdóttir, G. H., Rafnar, T., Thorlacius, S., Jonsson, E., Eyfjord, J. E. et al. (2007). Prostate cancer

progression and survival in BRCA2 mutation carriers. *Journal of the National Cancer Institute* **99** 929–935.

Tseng, P. (2001). Convergence of a block coordinate descent method for non-differentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Comput.* **23** 1661–1674. MR2839543

Vogel, D. and Fried, R. (2011). Elliptical graphical modelling. *Biometrika* **98** 935–951. MR2860334

Voorman, A., Shojaie, A. and Witten, D. (2014). Graph estimation with joint additive models. *Biometrika* **101** 85–101. MR3180659

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873 (2011f:62084)

Wang, H., Sun, D., Ji, P., Mohler, J. and Zhu, L. (2008). An AR-Skp2 pathway for proliferation of androgen-dependent prostate-cancer cells. *Journal of Cell Science* **121** 2578–2587.

Wang, Z., Gao, D., Fukushima, H., Inuzuka, H., Liu, P., Wan, L., Sarkar, F. H. and Wei, W. (2012). Skp2: a novel potential therapeutic target for prostate cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1825** 11–17.

Wu, Z., Cho, H., Hampton, G. M. and Theodorescu, D. (2009). Cdc6 and cyclin E2 are PTEN-regulated genes associated with human prostate cancer metastasis. *Neoplasia* **11** 66–76.

Yang, G., Ayala, G., De Marzo, A., Tian, W., Frolov, A., Wheeler, T. M., Thompson, T. C. and Harper, J. W. (2002). Elevated Skp2 protein expression in human prostate cancer: association with loss of the cyclin-dependent kinase inhibitor p27 and PTEN and with reduced recurrence-free survival. *Clinical Cancer Research* **8** 3419–3426.

Yang, E., Allen, G., Liu, Z. and Ravikumar, P. K. (2012). Graphical models via generalized linear models. In *Adv. Neural Inf. Process. Syst.* 1358–1366.

Yang, E., Ravikumar, P., Allen, G. I. and Liu, Z. (2013). On graphical models via univariate exponential family distributions. arXiv:1301.4183.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67. MR2212574

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94(10)** 19–35. MR2367824