Inference for High-dimensional Exponential Family Graphical Models

Jialei Wang
Department of Computer Science
University of Chicago
jialei@uchicago.edu

Mladen Kolar
Booth School of Business
University of Chicago
mkolar@chicagobooth.edu

Abstract

Probabilistic graphical models have been widely used to model complex systems and aid scientific discoveries. Most existing work on high-dimensional estimation of exponential family graphical models, including Gaussian and Ising models, is focused on consistent model selection. However, these results do not characterize uncertainty in the estimated structure and are of limited value to scientists who worry whether their findings will be reproducible and if the estimated edges are present in the model due to random chance. In this paper, we propose a novel estimator for edge parameters in an exponential family graphical models. We prove that the estimator is $\sqrt{n}$-consistent and asymptotically Normal. This result allows us to construct confidence intervals for edge parameters, as well as, hypothesis tests. We establish our results under conditions that are typically assumed in the literature for consistent estimation. However, we do not require that the estimator consistently recovers the graph structure. In particular, we prove that the asymptotic distribution of the estimator is robust to model selection mistakes and uniformly valid for a large number of data-generating processes. We illustrate validity of our estimator through extensive simulation studies.

1 Introduction

Probabilistic graphical models [Lauritzen, 1996] have been widely used to explore complex system and aid scientific discovery in areas ranging from biology and neuroscience to financial modeling and social media analysis. An undirected graphical model consists of a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of vertices and E is the set of edges, and a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$ that is Markov with respect to G. In particular, we have that X_a and X_b are conditionally independent given $X_{\backslash ab} = \{X_c \mid c \in \{1, \ldots, p\} \backslash \{a, b\}\}$ if and only if $(a, b) \notin E$. One of the central questions in high-dimensional statistics is estimation of the undirected graph G given n independent realizations of X, as well as quantifying uncertainty of the estimator.

We focus on a class of pairwise exponential family graphical models where the node conditional distribution of X_a given $X_{\backslash a} = \{X_c \mid c \in V \backslash a\}$ is specified by an exponential family

$$\log \mathbb{P}\left(X_a \mid X_{\backslash a}; \theta^*\right) = \Psi_a(X_a)\Big(\theta^*_{aa} + \sum_{b \in N(a)} \theta^*_{ab} \Psi_b(X_b)\Big) + C_a(X_a) - \bar{A}\Big(\theta^*_{aa} + \sum_{b \in N(a)} \theta^*_{ab} \Psi_b(X_b)\Big)$$

where $\{\Psi_a(\cdot)\}_{a \in V}$ are sufficient statistics, $C_a(X_a)$ is the base measure,

$$\bar{A}(t) = \log \int \exp\Big(\Psi_a(X_a) \cdot t + C_a(X_a)\Big) dX_a \qquad (1)$$

is the log-partition function, and $N(a) = \{b \in V \mid (a, b) \in E\}$ are neighbors of the node a in the graph G. These conditional distributions specify the following unique joint distribution [Yang et al., 2015]:

$$\begin{aligned} \log \mathbb{P}(X; \theta^*) = &\sum_{a \in V} \theta^*_{aa} \Psi_a(X_a) \\ &+ \sum_{(a,b) \in E} \theta^*_{ab} \Psi_a(X_a) \Psi_b(X_b) \\ &+ \sum_{a \in V} C_a(X_a) - A(\theta^*), \end{aligned} \qquad (2)$$

where $A(\theta^*)$ is the log partition function for the joint model. The model (2) considered here is general and includes Gaussian [Meinshausen and Bühlmann, 2006], Ising [Ravikumar et al., 2010] as special case. Given n independent observations $x_1, \ldots, x_n$ from the model in (2), we construct a $\sqrt{n}$ consistent estimator of a parameter θ^*_{ab} and

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

show that it is asymptotically normal. Based on this result, we perform (asymptotic) inference for coefficients θ^*. In particular, we construct valid confidence intervals for parameters in the model with nominal coverage and propose statistical tests with nominal size to test existence of edges in the graphical model. Our inference results are uniformly valid over a large class of data generating procedures and are robust to model selection mistakes, which commonly occur in ultra-high dimensional setting. Our results are of fundamental importance to scientists who are interested in uncertainty associated with point estimates. For example, given a point estimate $\widehat{\theta}$ of θ^*, a scientist does not know if an edge is present in the estimated model due to random fluctuation or there is indeed a statistically significant conditional dependence between two nodes. Therefore, our results complement existing results in the literature, which are focused on consistent model selection and parameter recovery, as we review below.

Related work. Our work contributes to two areas. First, we contribute to the growing literature on graphical model selection in high-dimensions. A lot of work has been done under the assumption that $X \sim N(0, \Sigma)$, in which case the edge set E of the graph G is encoded by the non-zero elements of the precision matrix Ω [Meinshausen and Bühlmann, 2006, Yuan and Lin, 2007, Rothman et al., 2008, Friedman et al., 2008, d'Aspremont et al., 2008, Fan et al., 2009, Lam and Fan, 2009, Yuan, 2010, Cai et al., 2011, Liu and Wang, 2012, Zhao and Liu, 2014]. Learning structure of the Ising model based on the penalized pseudo-likelihood was studied in [Höfling and Tibshirani, 2009, Ravikumar et al., 2010, Xue et al., 2012]. More recently, [Yang et al., 2015] studied estimation of graphical models under the assumption that the node conditional distribution belongs to an exponential family distribution. Such node conditional distribution includes many standard distributions, including Bernoulli, Gaussian, Poisson and exponential. Furthermore, there exist joint distributions, consistent with these node conditional distributions, that include large number of familiar graphical models, including Gaussian graphical model, Ising model and mixed graphical models [Lee and Hastie, 2012, Chen et al., 2013a, Cheng et al., 2013, Yang et al., 2014, 2013, 2015, Yuan et al., 2013, Guo et al., 2011]. In our paper, we construct a novel $\sqrt{n}$ consistent estimator of a parameter corresponding to a particular edge in a pairwise exponential family graphical model. This is a first procedure that can obtain parametric rate of convergence for an edge parameter in an exponential family graphical model.

Second, we contribute to the literature on high-dimensional inference. Recently, there has been a lot of interest on performing valid statistical inference in the high-dimensional setting. [Zhang and Zhang, 2013, Belloni et al., 2013a,b, van de Geer et al., 2014, Javanmard and Montanari, 2014, 2013, Ning and Liu, 2014a,b, Farrell, 2013] developed methods for construction of confidence intervals for low dimensional parameters in high-dimensional linear and generalized linear models, as well as hypothesis tests. These methods construct honest, uniformly valid confidence intervals and hypothesis test based on the ℓ_1 penalized estimator in the first stage. Similar results were obtained in the context of ℓ_1 penalized least absolute deviation and quantile regression [Belloni et al., 2013c,d]. [Lockhart et al., 2014] study significance of the input variables that enter the model along the lasso path. [Lee et al., 2013, Taylor et al., 2014] perform post-selection inference conditional on the selected model. [Liu, 2013, Ren et al., 2013, Chen et al., 2013b] construct $\sqrt{n}$ consistent estimators for elements of the precision matrix Ω under a Gaussian assumption. We contribute to the literature by demonstrating how to construct estimators that are robust and uniformly valid under more general distributional assumptions.

Notation. We use $[n]$ to denote the set $\{1, \ldots, n\}$. For a vector $a \in \mathbb{R}^n$, let $S(a) = \{j : a_j \neq 0\}$ be the support set. Let $\Lambda_{\max}(A)$ denote the maximum eigenvalue of matrix A. Let $A \circ B$ denote the Hadamard product of A and B. We use $N(\mu, \sigma^2)$ to denote the normal distribution with mean μ and variance σ^2. We use $a_n \lesssim b_n$ to denote that $a_n \leq C b_n$ holds for all large enough n, with some finite positive constant C, and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$. We use $\rightarrow_{\mathcal{D}}$ to denote convergence in distribution.

2 Methodology

In this section, we outline our double selection procedure for estimating θ^*_{ab} and making valid inference about this parameter. To simplify the presentation, we assume that $\Psi_a(X_a) = X_a$ and $\Psi_{ab}(X_a, X_b) = X_a X_b$. Let $x_1, \ldots, x_n$ be n independent samples from (2). With out loss of generality, assume we are interested in the edge between a and b, where $a, b \in [p]$, and $a < b$. Let $\theta \in \mathbb{R}^{2p-1}$ be the vector such that

$$\theta = (\theta_{ab}, \underbrace{\theta_{1a}, \ldots, \theta_{pa}}_{\text{no index for } b}, \underbrace{\theta_{1b}, \ldots, \theta_{pb}}_{\text{no index for } a})^T.$$

Our procedure consists of three steps. In the first step, we construct a pilot estimator $\widetilde{\theta}$ of θ^* by minimizing the penalized composite likelihood (see (5) below). The goal of this step is to find a parameter θ that solves the population set of estimating equations

$$\mathbb{E}[\nabla L_n(\theta)] = 0, \tag{3}$$

where $L_n(\theta) = \frac{1}{2}(L_n^a(\theta) + L_n^b(\theta))$ is the composite likelihood and $L_n^a(\theta)$ is the negative conditional pseudo-likelihood for node $a \in V$ given by

$$L_n^a(\theta) = -\frac{1}{n}\sum_{i\in[n]} x_{ia}\left(\theta_{aa} + \sum_{b\in N(a)} \theta_{ab}x_{ib}\right) + C_a(x_{ia}) - \bar{A}\left(\theta_{aa} + \sum_{b\in N(a)} \theta_{ab}x_{ib}\right).$$

We obtain $\widetilde{\theta}$ by first finding a minimizer of the penalized composite likelihood

$$\widehat{\theta} = \arg\min_\theta L_n(\theta) + \lambda_1\|\theta\|_1$$

and then refitting on the estimated support

$$\widetilde{\theta} = \arg\min_\theta L_n(\theta) \quad \text{subject to} \quad S(\widetilde{\theta}) \subseteq S(\widehat{\theta}). \tag{4}$$

Since $\widetilde{\theta}$ is obtained via a model selection procedure, it is irregular and its asymptotic distribution cannot be estimated [Leeb and Pötscher, 2007, Pötscher, 2009]. In order to make the estimator regular, in steps 2 and 3 we construct $\check{\theta}$ that is robust against mistakes in estimation of the nuisance component[1] of the vector θ^*. In particular, we find θ, which, in addition to (4), also solves for

$$\frac{\partial}{\partial\theta_{\setminus ab}}\mathbb{E}[\nabla L_n(\theta)]\Big|_{\theta_{\setminus ab}=\theta^*_{\setminus ab}} = 0. \tag{5}$$

The relation (6) states that the estimator needs to be insensitive with respect to first order perturbations of the nuisance. The idea of creating an estimator that is robust to perturbations of nuisance have been recently used in [Belloni et al., 2013a] and [Belloni et al., 2013d], however, the approach goes back to work of [Neyman, 1959]. In the remainder of the section, we provide details to the steps two and three.

In order to facilitate exposition, we introduce additional notation. Let $\mathcal{X}_{ia} \in \mathbb{R}^{2p-1}$ be a vector with p non-zero elements: the element in the first position, $\mathcal{X}_{ia(1)} = x_{ib}$, and the elements in position 2 to position b, $\mathcal{X}_{ia(j)} = x_{i(j-1)}$, except the position $a+1$, where $\mathcal{X}_{ia(a+1)} = 1$; from position $b+1$ to position p, $\mathcal{X}_{ia(j)} = x_{i(j)}$, the elements in position $p+1$ to position $2p-1$ are all zeros. That is

$$\mathcal{X}_{ia} = (x_{ib}, x_{i1}, \ldots, x_{i(a-1)}, 1, \ldots, x_{ip}, \underbrace{0, \ldots, 0}_{p+1 \text{ to } 2p-1})^T.$$

Thus

$$L_n^a(\theta) = -\frac{1}{n}\sum_{i\in[n]} x_{ia}\left(\mathcal{X}_{ia}^T\theta\right) + C_a(x_{ia}) - \bar{A}\left(\mathcal{X}_{ia}^T\theta\right).$$

[1]Since we are interested in inference for θ^*_{ab}, we treat the rest of the vector as a nuisance.

similarly, we can define $\mathcal{X}_{ib} \in \mathbb{R}^{2p-1}$ as

$$\mathcal{X}_{ib} = (x_{ia}, \underbrace{0, \ldots, 0}_{2 \text{ to } p}, x_{i1}, \ldots, x_{i(b-1)}, 1, \ldots, x_{ip})^T.$$

Let $Q_n(\theta)$ be the Hessian of L_n evaluated at θ, with elements

$$[Q_n(\theta)]_{(i)(j)} = \frac{\partial^2 L_n(\theta)}{\partial\theta_i\partial\theta_j}.$$

With this notation, we are ready to give details of step 2 and 3 of our estimation procedure for the edge parameter θ^*_{ab}. In the second step, we find $\widehat{\gamma}_{ab} \in \mathbb{R}^{2p-1}$ by minimizing the following ℓ_1 penalized problem

$$\min_\gamma \frac{1}{2}\gamma^T Q_n(\widetilde{\theta})\gamma - e_{ab}^T Q_n(\widetilde{\theta})\gamma + \lambda_2\|\widehat{\Gamma}\gamma\|_1 \quad \text{subject to} \quad \gamma_{ab} = 0, \tag{6}$$

where $\widehat{\Gamma} \in \mathbb{R}^{2p-1\times 2p-1}$ is a diagonal weighting matrix. Each diagonal element $\widehat{\Gamma}^2_{ab}$ serves as an estimate of the variance of the score vector. Here we remark that the choice of $\widehat{\Gamma}$ allows us to choose the penalty parameter λ_2 is a way that does not depend on unknown parameters of the problem.

Finally, in the third step, we obtain a consistent, asymptotically normal estimator of θ^*_{ab} by minimizing the following restricted optimization program

$$\check{\theta} = \arg\min_\theta L_n(\theta) \quad \text{subject to} \quad S(\check{\theta}) \subseteq \widetilde{S} \tag{7}$$

where $\widetilde{S} = S(\widetilde{\theta}) \cup S(\widehat{\gamma}) \cup \{ab\}$. We will show that

$$\widehat{\sigma}^{-1}_{n,ab}\sqrt{n}\left(\check{\theta}_{ab} - \theta^*_{ab}\right) \to_{\mathcal{D}} N(0,1) \quad \text{where} \quad \widehat{\sigma}^2_{n,ab} = \left[\left(Q_{n,\widetilde{S}\widetilde{S}}(\widetilde{\theta})\right)^{-1}\right]_{ab,ab}. \tag{8}$$

Let z_α be such that $\mathbb{P}[N(0,1) \le z_\alpha] = \alpha$. Based on the above result, we construct a $1-\alpha$ confidence interval for θ^*_{ab} as

$$\left[\check{\theta}_{ab} + \widehat{\sigma}_n z_{\alpha/2}/\sqrt{n}, \check{\theta}_{ab} + \widehat{\sigma}_n z_{1-\alpha/2}/\sqrt{n}\right]. \tag{9}$$

This confidence interval is uniformly valid over a large number of data generating processes and does not rely on consistent model selection or β-min condition that is commonly assumed for proving sparsistency [Wainwright, 2009].

It is not immediately obvious why should this procedure work. Therefore, we provide a heuristic argument here. Let $\check{Q} = Q_n(\check{\theta})$. The first order optimality condition give us $[\nabla L_n(\check{\theta})]_{\widetilde{S}} = 0$ and

$$\begin{aligned}
[\nabla L_n(\theta^*)]_{\widetilde{S}} &= [\nabla L_n(\theta^*) - \nabla L_n(\check{\theta})]_{\widetilde{S}} \\
&= \Big[\underbrace{\left(\int_0^1 Q_n(t\check{\theta} + (1-t)\theta^*)\mathrm{d}t\right)}_{\bar{Q}:=\bar{Q}(\check{\theta})}\Big]_{\widetilde{S}}(\check{\theta} - \theta)_{\widetilde{S}} \\
&= \check{Q}_{\widetilde{S}}(\check{\theta} - \theta)_{\widetilde{S}} + \left(\bar{Q} - \check{Q}\right)_{\widetilde{S}}(\check{\theta} - \theta)_{\widetilde{S}}.
\end{aligned} \tag{10}$$

Let $\check{\gamma} = \gamma(\check{\theta}, \widetilde{S}) \in \mathbb{R}^{|\widetilde{S}|}$ be the minimizer of

$$\min_{\gamma} \frac{1}{2}\gamma^T \check{Q}_{\widetilde{S}}\gamma - e_{ab}^T \check{Q}_{\widetilde{S}}\gamma \quad \text{subject to} \quad \gamma_{ab} = 0 \tag{11}$$

and $\sigma_{n,ab}^2 = \check{Q}_{ab,ab} - 2e_{ab}^T \check{Q}_{\widetilde{S}}\check{\gamma} + \check{\gamma}^T \check{Q}_{\widetilde{S}}\check{\gamma}$. Let $\check{c} = c(\check{\theta}, \widetilde{S}) \in \mathbb{R}^{|\widetilde{S}|}$ be the contrast vector with components defined as $\check{c}_j = -\check{\gamma}_j$ for $j \neq ab$, and $\check{c}_j = 1$ for $j = ab$. In a similar way, we define $c(\theta^*, \widetilde{S})$ with $Q_n(\theta^*)_{\widetilde{S}}$ used in (12) to obtain $\gamma(\theta^*, \widetilde{S})$. Multiplying (11) by $\check{c}$ and rearranging terms, we have

$$\begin{aligned} &c(\theta^*, \widetilde{S})^T \left[\nabla L_n(\theta^*)\right]_{\widetilde{S}} + \Delta^{(2)}(\check{\theta}) \\ &= \sigma_{n,ab}^2 \left(\check{\theta}_{ab} - \theta_{ab}\right) + \Delta^{(1)}(\check{\theta}) \end{aligned}$$

where $\Delta^{(1)}(\check{\theta}) = \check{c}^T(\bar{Q} - \check{Q})_{\widetilde{S}}(\check{\theta} - \theta)_{\widetilde{S}}$ and $\Delta^{(2)}(\check{\theta}) = \left(\check{c} - c(\theta^*, \widetilde{S})\right)^T \left[\nabla L_n(\theta^*)\right]_{\widetilde{S}}$. Using the properties of the refitted estimator, we will prove that $\sqrt{n}\Delta^{(1)}(\check{\theta}) \to_{\mathcal{P}} 0$ and $\sqrt{n}\Delta^{(2)}(\check{\theta}) \to_{\mathcal{P}} 0$ uniformly for all $\check{\theta}$ in a neighborhood around θ^*. Therefore, our heuristic argument is almost done as $\sigma_{n,ab}^2$ converges in probability to a quantity that is bounded away from zero and $\sqrt{n}c(\theta^*, \widetilde{S})^T \left[\nabla L_n(\theta^*)\right]_{\widetilde{S}}$ has a limiting normal distribution. All that we still need to show is that σ_n^2 in (9) consistently estimates the asymptotic variance of $\sqrt{n}(\check{\theta}_{ab} - \theta_{ab}^*)$. This heuristic argument is made precise in Theorem 8.

It should be clear from the above argument that we did not require that $S(\theta^*) \subseteq \widetilde{S}$. Our results only require that the refitted composite likelihood estimator consistently estimates θ^* in the ℓ_2 norm. However, this can be established under mild conditions [Negahban et al., 2012].

3 Theoretical Properties

In this section, we outline main theoretical properties of our procedure. We start by providing high-level conditions that allow us to establish properties of each step in our procedure. Let $Q^* = \mathbb{E}[W_{ia}(\theta^*)^2 \mathcal{X}_{ia}\mathcal{X}_{ia}^T + W_{ib}(\theta^*)^2 \mathcal{X}_{ib}\mathcal{X}_{ib}^T]$ be the population version of the Fisher information matrix, where $W_{ia}(\theta^*)^2 = \bar{A}''(\mathcal{X}_{ia}^T\theta^*)$. Let $U_n = (2n)^{-1}\sum_{i\in[n]}(\mathcal{X}_{ia}\mathcal{X}_{ia}^T + \mathcal{X}_{ib}\mathcal{X}_{ib}^T)$ and $U^* = \mathbb{E}[U_n]$ be the sample and population versions of the covariance matrix, respectively. Let $s = |S(\theta^*)|$ be the size of true support. We assume the following two regularity conditions on $Q_n(\theta^*)$ and U_n.

Assumption 1. *(Restricted eigenvalue) There exists a constant $\kappa > 0$, such that for any Δ satisfying $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$ with $S = \text{support}(\theta^*)$, we have $\Delta^T Q_n \Delta \geq \kappa\|\Delta\|_2^2$.*

Let $\phi_A(s) = \sup\left\{\frac{\Delta^T A\Delta}{\|\Delta\|_2^2} \mid \|\Delta\|_0 \leq s\right\}$ denote the maximum s-sparse eigenvalue of a matrix A.

Assumption 2. *(Maximum sparse eigenvalue) There exists $\phi_{\max}$, such that $\phi_{U_n}(s) = \phi_{\max} < \infty$.*

Above conditions are required to establish estimation consistency and sparsity control in the first stage. Assumption A1 is weaker than the Incoherence condition in [Yang et al., 2012], which is needed to ensure model selection consistency. Assumption A2 is required to show that with high-probability $|\widetilde{S}| \leq C|S(\theta^*)|$, that is, the size of the selected support in the third stage is not too large. As shown in the appendix, by assuming the sparse eigenvalue conditions on population Fisher information matrix and convariance matrix, these conditions can be shown to hold with high-probability.

Similar to [Yang et al., 2012], we need the following conditions on the exponential family graphical model to obtain tail bounds on X.

Assumption 3 (Moment). *There exist constants κ_1 and κ_2 such that $\max_{a\in V}\mathbb{E}[X_a] \leq \kappa_1$ and $\max_{a\in V}\mathbb{E}[X_a^2] \leq \kappa_2$.*

Assumption 4 (Log-partition function of joint distribution). *There exists a constant κ_A, such that $\sup_{u:|u|\leq 1}(\partial^2/\partial^2\theta_{ab})A\left(\theta_{ab}^* + u\right) \leq \kappa_A$.*

Assumption 5 (Log-partition function of node–conditional distribution). *There exist constants $\kappa_3 \max_a \frac{9}{2}\|\theta_a^*\|_2$ and $\kappa_4 \in [0, 1/4]$ such that $\max\{|\bar{A}''(\kappa_3 \log p)|, |\bar{A}'''(\kappa_3 \log p)|\} \leq n^{\kappa_4}$.*

With the above assumptions, we have the following result that characterizes the performance of the first stage estimator.

Theorem 6. *Suppose that assumptions A1-A5 hold and that the penalty parameter in the first stage satisfies $\lambda_1 = 2C_1\sqrt{\frac{\log p}{n}}$ for some constant C_1 that does not depend on (n, p, s). Then there exists a constant C_2, such that with probability at least $1 - \mathcal{O}(p^{-1})$, we have*

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{C_2}{\kappa}\sqrt{\frac{s\log p}{n}}, \quad \|\widehat{\theta} - \theta^*\|_1 \leq \frac{4sC_2}{\kappa}\sqrt{\frac{\log p}{n}},$$
$$(\widetilde{\theta} - \theta^*)^T Q_n(\theta^*)(\widetilde{\theta} - \theta^*) \leq \frac{256 sC_2 \log p}{\kappa n}$$
$$|S(\widetilde{\theta})| \leq s\frac{C_2^2 \min_{m\in\mathcal{M}}\phi_{U_n}(m)}{64\kappa^2},$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > 2sC_1^2\phi_{U_n}(m)/(64\kappa^2)\}$.

This theorem establishes two results. First, the ℓ_2 norm convergence result for the stage one estimator. Second, it establishes that the size of the support of $\widetilde{\theta}$ is not much larger than the size of the true support θ^*.

Our next result establishes convergence results for the second stage. Let γ_{ab}^* be the minimizer of

$$\min_{\gamma} \frac{1}{2}\gamma^T Q^*\gamma - e_{ab}^T Q^*\gamma \quad \text{subject to} \quad \gamma_{ab} = 0.$$

Under the sparsity assumption on θ^*, we have that the Markov blanket of (X_a, X_b) is small, hence γ^*_{ab} is sparse. The following theorem establish the estimation error bound on γ^*.

Theorem 7. *Suppose that assumptions A1-A5 hold and that the penalty parameter in the second stage satisfies $\lambda_2 = 2C_1\sqrt{\frac{\log p}{n}}$. Then with probability at least $1 - \mathcal{O}(p^{-1})$,*

$$(\widehat{\gamma} - \gamma^*)^T Q_n(\widetilde{\theta})(\widehat{\gamma} - \gamma^*) \leq C\frac{s\log p}{\kappa n}$$

$$|\text{support}(\widehat{\gamma})| \leq |\text{support}(\gamma^*)|\frac{C_1^2 \min_{m\in\mathcal{M}} \phi_{U_n}(m)}{9\kappa^2}$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > 2sC_1^2\phi_{U_n}(m)/(9\kappa^2)\}$.

We will use results of Theorem 6 and 7 to prove our main result given in the following theorem.

Theorem 8. ***(Asymptotic Normality)*** *Suppose that conditions of Theorem 6 and 7 hold. The estimator $\check{\theta}_{ab}$ admits the following decomposition*

$$\widehat{\sigma}^{-1}_{n,ab}\sqrt{n}(\check{\theta}_{ab} - \theta^*_{ab}) = N_{ab} + o_P(1)$$

where $N_{ab} \to_{\mathcal{D}} N(0, 1)$ and $\widehat{\sigma}_{n,ab}$ is defined in (9).

Note that the result holds for a wide range of data generating processes and does not require perfect model selection. Our estimator is regular and robust to model selection mistakes. We will illustrate these properties in simulation studies where we also compare with a naive estimator that assumes perfect model selection.

4 Simulation Studies

In this section we perform extensive simulation studies to illustrate finite sample performance of our procedure. We demonstrate that the proposed double selection approach can be used to construct valid confidence intervals for various exponential family graphical models (including Gaussian, Ising and Poisson) on various graph structures (including Chain, Nearest Neighbor and Erdős-Rényi).

We construct three kinds of graph structures, described below:

Chain graph: we first randomly permute the nodes and then connect them in succession.

Nearest Neighbor graph: we follow the generating process described in [Li and Gui, 2006]. For each node, we draw a point uniformly at random in a unite square, then connect the nodes to its 4-Nearest Neighbors.

Erdős-Rényi graph: we follow the process in [Meinshausen and Bühlmann, 2006]. We generate a random graph with $2p$ edges and a constrain that the maximum degree cannot exceed 5.

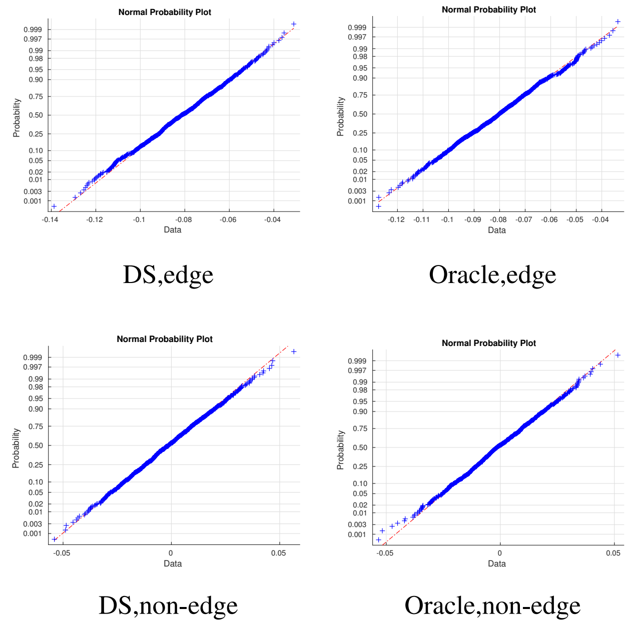


Figure 2: Normal probability plot of estimations on Poisson, chain graph, n = 200, p = 200.

We generate the following three examples in exponential family graphical models, with the detailed settings described blow:

Gaussian graphical model: where the node conditional distribution is

$$X_a|X_{\backslash a} = x_{\backslash a} \sim N\left(\sum_{(a,b)\in S^*} \frac{\theta^*_{ab}}{\theta^*_{bb}}x_b, 1\right).$$

We set $\theta^*_{ii}, \forall i \in [p]$ to 1, and for each edge $(a, b) \in S^*$, the strength θ^*_{ab} is drawn from uniformly random distribution in $[-0.5, 0.5]$.

Ising graphical model: where the node conditional distribution is

$$X_a|X_{\backslash a} = x_{\backslash a} \sim \text{Bern}\left(\frac{\exp(2X_a\sum_{(a,b)\in S^*}\theta^*_{ab}x_b)}{\exp(2X_a\sum_{(a,b)\in S^*}\theta^*_{ab}x_b) + 1}\right).$$

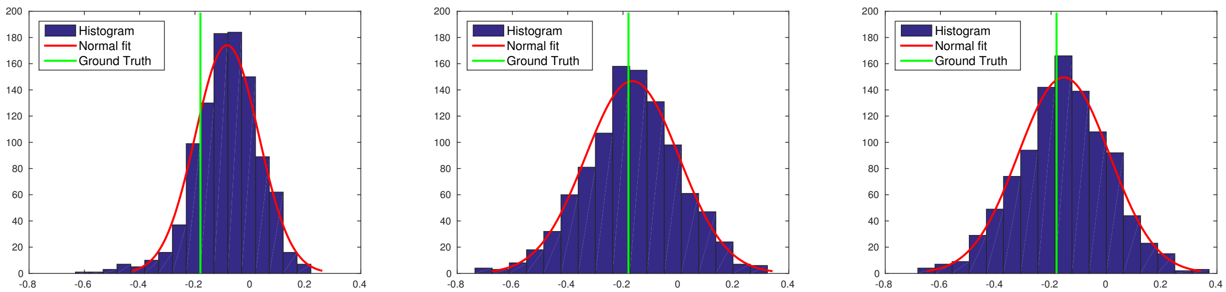
For each edge $(a, b) \in S^*$, we choose θ^*_{ab} uniformly in $[-1, 1]$.

Poisson graphical model: where the node conditional distribution is

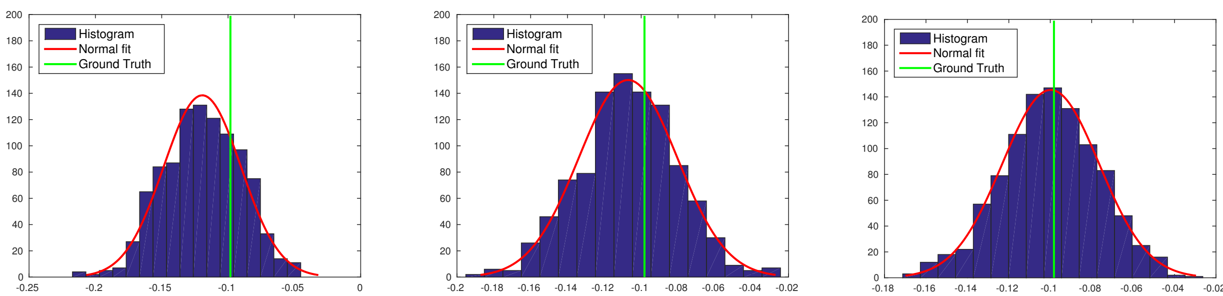
$$X_a|X_{\backslash a} = x_{\backslash a} \sim \text{Poisson}\left(\sum_{(a,b)\in S^*} \theta^*_{ab}x_b\right).$$

We set $\theta^*_{ii}, \forall i \in [p]$ as 2, and for each edge $(a, b) \in S^*$, the strength θ^*_{ab} is drawn uniformly in $[-0.2, 0]$.

Note that under the above settings, the true edge strength might be weak, which makes the consistent model selection unlikely. Except for the Gaussian graphical model



Naive selection | Double selection | Oracle estimation



Naive selection | Double selection | Oracle estimation

Figure 1: Comparison between Naive Selection, Double Selection and Oracle Estimator on a random edge of the graph with $n = 200, p = 200$. Top row: Ising models with Nearest Neighbor graph, Bottom row: Poisson models with Erdős-Rényi graph.

from which the data could be efficiently sampled, we use Markov Chain Monte Carlo sampling to generate the data. We solve the joint neighborhood selection procedure using proximal gradient methods, with λ_1 and λ_2 set to be a constant factor of $\sqrt{\log p/n}$ following our theories. For each setting, we first randomly draw a pair from S (edges) and a pair from S^c (non-edges) as the node pair we are interested in performing inference.

4.1 Asymptotic Normality

We first demonstrate the asymptotic normality of the proposed estimator. We run the inference procedure 1000 times and draw the histogram of $\check{\theta}_{ab}$ together with a probability density function of a normal distribution and the true value θ^*_{ab} as a reference. We compare the double selection procedure with the post naive selection, which uses lasso to select the edges and perform post selection inference, and the oracle estimator, which knows the graph structure S^* and performs maximum likelihood estimation on S^*. Figure 1 shows a case for Ising and Poisson graphical models[2]. We can clearly see that if we just perform post naive selection inference, the estimator is significantly biased which makes the constructed confidence intervals invalid. The post double selection procedure is robust to model selection mistakes and produces unbiased confidence intervals with correct coverage. Compared to the oracle, the sampling variance of the estimator is slightly larger, which is the price we pay for not knowing the true graph structure. Finally, Figure 2 shows the normal probability plot for the estimation of random pairs on a Poisson model with chain graph structure. Again, the normal probability plot demonstrate the estimations obtained via the proposed post-double selection procedure are comparable to those obtained from the oracle procedure.

4.2 Confidence Intervals

We also examine the performance of the constructed confidence intervals. We construct a 95% confidence interval based on (10). We measure the empirical frequency that the constructed confidence interval covers the true parameter, and the average width of the confidence intervals. Suppose we are interested in θ^*_{ab} and let $\widehat{\mathrm{CI}}_t(\theta_{ab})$ denote the constructed interval at t-th trial. Then the average coverage and average length of the confidence interval will be

$$\text{Avgcov} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\{\theta^*_{ab} \in \widehat{\mathrm{CI}}_t(\theta_{ab})\}}$$

$$\text{Avglen} = \frac{1}{T}\sum_{t=1}^{T} |\widehat{\mathrm{CI}}_t(\theta_{ab})|.$$

As a reference, we also compared with the oracle proce-

[2]Please refer to appendix for more results.

Table 1: Coverage and length of constructed confidence intervals for Gaussian graphical models

Graph(n,p)	DS, $(a,b) \in S$		DS, $(a,b) \in S^c$		Oracle, $(a,b) \in S$		Oracle, $(a,b) \in S^c$	
	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen
Chain(200,100)	0.998	0.363	0.975	0.321	0.994	0.318	0.969	0.300
ER(200,100)	0.944	0.267	0.915	0.257	0.903	0.219	0.919	0.226
NN(200,100)	0.983	0.340	0.988	0.397	0.967	0.295	0.995	0.347
Chain(200,200)	0.986	0.362	0.976	0.353	0.972	0.300	0.985	0.310
ER(200,200)	0.935	0.264	0.921	0.266	0.883	0.210	0.918	0.222
NN(200,200)	0.987	0.402	0.997	0.415	0.989	0.316	0.999	0.347
Chain(400,200)	0.971	0.229	0.969	0.227	0.981	0.209	0.982	0.216
ER(400,200)	0.916	0.163	0.882	0.165	0.854	0.146	0.905	0.154
NN(400,200)	0.982	0.248	0.987	0.260	0.985	0.219	0.999	0.239

Table 2: Coverage and length of constructed confidence intervals for Ising graphical models

Graph(n,p)	DS, $(a,b) \in S$		DS, $(a,b) \in S^c$		Oracle, $(a,b) \in S$		Oracle, $(a,b) \in S^c$	
	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen
Chain(200,100)	0.935	0.488	0.992	0.393	0.954	0.421	0.981	0.336
ER(200,100)	0.966	0.564	0.971	0.524	0.974	0.418	0.975	0.442
NN(200,100)	0.962	0.492	0.993	0.473	0.958	0.468	0.988	0.473
Chain(200,200)	0.966	0.500	0.983	0.378	0.975	0.374	0.982	0.353
ER(200,200)	0.978	0.745	0.983	0.587	0.979	0.507	0.988	0.455
NN(200,200)	0.986	0.830	0.982	0.425	0.976	0.754	0.968	0.341
Chain(400,200)	0.964	0.341	0.981	0.262	0.978	0.262	0.969	0.247
ER(400,200)	0.965	0.522	0.977	0.402	0.981	0.350	0.984	0.313
NN(400,200)	0.968	0.543	0.984	0.285	0.968	0.512	0.971	0.238

dure, which performs classical statistical inference with known true structure. Table 1,2,3 report the Avgcov and Avglen on Gaussian, Ising, and Poisson graphical models, respectively. We make the following observations:

i) When looking at the coverage, both the post double selection procedure and oracle inference performs reasonably well for all kinds of graphical models and graph structures, which verifies the theories of asymptotic normality. As the number of samples n increase, the coverage tends to be closer to 95%.

ii) When looking at the width, we found that generally speaking the confidence intervals produced by post double selection are just slightly wider than the ones in oracle inference, which is considered to be the asymptotically efficient estimators. This validates the power of proposed double selection procedure.

iii) When comparing the width across different graph structures, on Ising and Poisson models: the Nearest Neighbor and Erdős-Rényi graphs usually produce wider confidence intervals than chain graphs, which might because that chain graph is relatively easier for graph recovery.

4.3 Illustration

We also give illustrations to demonstrate constructed confidence intervals according to our procedure. We perform post double selection inference for every edge in the graph and then plot the 95% confidence intervals. Figure 3 provides one realization of confidence intervals of the whole node paris (most elements in S^c were omitted for better visualization), for an Ising graphical model with chain graph structure. We can see that the intervals are quite reasonable and most of them trap the true parameter.

5 Conclusion

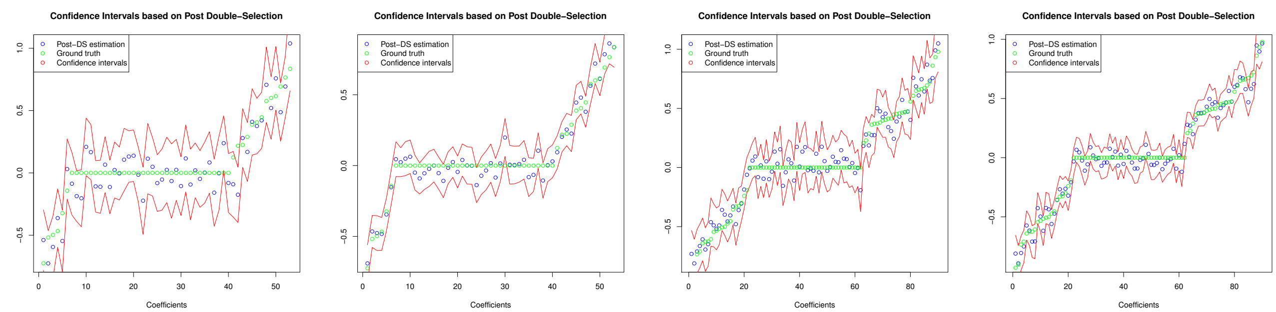
We develop a new robust estimation procedure for an edge parameter in an exponential family graphical model. Our estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal, which allows us to perform statistical inference. This is very important problem with huge practical implications. Our paper timely fills a gap in the literature that has so far been focused on point estimation. Our theoretical results are illustrated through simulations on commonly used types of probabilistic graphical models, where the node-conditional distribution is Gaussian, Bernoulli and Poisson.

References

S. L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford

Table 3: Coverage and length of constructed confidence intervals for Poisson graphical models

Graph(n,p)	DS, $(a,b) \in S$		DS, $(a,b) \in S^c$		Oracle, $(a,b) \in S$		Oracle, $(a,b) \in S^c$	
	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen	Avgcov	Avglen
Chain(200,100)	0.973	0.074	0.921	0.085	0.928	0.056	0.939	0.070
ER(200,100)	0.964	0.178	0.952	0.084	0.964	0.143	0.945	0.060
NN(200,100)	0.950	0.062	0.936	0.067	0.881	0.046	0.927	0.054
Chain(200,200)	0.960	0.067	0.928	0.065	0.903	0.052	0.914	0.053
ER(200,200)	0.965	0.118	0.937	0.088	0.957	0.096	0.928	0.072
NN(200,200)	0.952	0.100	0.951	0.093	0.944	0.080	0.945	0.075
Chain(400,200)	0.973	0.048	0.938	0.048	0.890	0.037	0.918	0.037
ER(400,200)	0.976	0.085	0.946	0.068	0.951	0.068	0.937	0.051
NN(400,200)	0.961	0.072	0.930	0.068	0.946	0.056	0.941	0.052



p = 20, n = 100 p = 20, n = 400 p = 50, n = 200 p = 50, n = 400

Figure 3: Illustration of constructed confidence interval on Ising graphical model with chain graph, the true parameters are in green and the post double selection estimators are in blue. We removed most elements in S^c for better visualization.

University Press, New York, 1996. ISBN 0-19-852219-3. Oxford Science Publications.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 2015.

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

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

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

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

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, 30(1):56–66, 2008.

J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat.*, 3(2): 521–541, 2009.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.*, 37: 4254–4278, 2009.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286, 2010.

T. T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *ArXiv e-prints, arXiv:1209.2437*, September 2012.

T. Zhao and H. Liu. Calibrated precision matrix estimation for high dimensional elliptical distributions. *IEEE Trans. Inf. Theory*, pages 1–1, 2014.

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

L. Xue, H. Zou, and T. Ca. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *Ann. Stat.*, 40(3):1403–1429, 2012.

J. D. Lee and T. J. Hastie. Learning mixed graphical models. *ArXiv e-prints, arXiv:1205.5012*, May 2012.

S. Chen, D. M. Witten, and A. Shojaie. Selection and estimation for mixed graphical models. *ArXiv e-prints, arXiv:1311.0085*, November 2013a.

J. Cheng, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *ArXiv e-prints, arXiv:1304.2810*, April 2013.

E. Yang, Y. Baker, P. Ravikumar, G. I. Allen, and Z. Liu. Mixed graphical models via exponential families. In *Proc. 17th Int. Conf, Artif. Intel. Stat.*, pages 1042–1050, 2014.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On poisson graphical models. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1718–1726. Curran Associates, Inc., 2013.

X.-T. Yuan, P. Li, and T. Zhang. Learning pairwise graphical models with nonlinear sufficient statistics. *ArXiv e-prints, arXiv:1311.5479*, November 2013.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1): 1–15, 2011.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. B*, 76(1):217–242, Jul 2013.

A. Belloni, V. Chernozhukov, and C. B. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.*, 81(2):608–650, Nov 2013a.

A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969*, 2013b.

S. A. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.*, 42(3):1166–1202, Jun 2014.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(Oct):2869–2909, 2014.

A. Javanmard and A. Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1311.0274*, November 2013.

Y. Ning and H. Liu. Sparc: Optimal estimation and asymptotic inference under semiparametric sparsity. *ArXiv e-prints, arXiv:1412.2295*, 2014a.

Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *ArXiv e-prints, arXiv:1412.8765*, 2014b.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint arXiv:1309.4686*, September 2013.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post selection inference for lad regression models. *arXiv preprint arXiv:1304.0282*, 2013c.

A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional approximately sparse quantile regression models. *arXiv preprint arXiv:1312.7186*, December 2013d.

R. Lockhart, J. E. Taylor, R. J. Tibshirani, and R. J. Tibshirani. A significance test for the lasso. *Ann. Stat.*, 42(2): 413–468, 2014.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference with the lasso. *ArXiv e-prints, arXiv:1311.6238*, November 2013.

J. E. Taylor, R. Lockhart, R. J. Tibshirani, and R. J. Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, January 2014.

W. Liu. Gaussian graphical model estimation with false discovery rate control. *Ann. Stat.*, 41(6):2948–2978, 2013.

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013.

M. Chen, Z. Ren, H. Zhao, and H. H. Zhou. Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *arXiv preprint arXiv:1309.5923*, 2013b.

H. Leeb and B. M. Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econ. Theory*, 24(02):338–376, Nov 2007.

B. M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.

J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. *Probability and statistics*, 57:213, 1959.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202, 2009.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Stat. Sci.*, 27(4):538–557, 2012.

E. Yang, G. I. Allen, Z. Liu, and P. Ravikumar. Graphical models via generalized linear models. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1358–1366. Curran Associates, Inc., 2012.

H. Li and J. Gui. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2006.