Generalized Direct Change Estimation in Ising Model Structure

Farideh Fazayeli FARIDEH@CS.UMN.EDU
Arindam Banerjee BANERJEE@CS.UMN.EDU
Department of Computer Science & Engineering, University of Minnesota, Twin Cities

Abstract

We consider the problem of estimating change in the dependency structure of two p-dimensional Ising models, based on respectively n_1 and n_2 samples drawn from the models. The change is assumed to be structured, e.g., sparse, block sparse, node-perturbed sparse, etc., such that it can be characterized by a suitable (atomic) norm. We present and analyze a norm-regularized estimator for directly estimating the change in structure, without having to estimate the structures of the individual Ising models. The estimator can work with any norm, and can be generalized to other graphical models under mild assumptions. We show that only one set of samples, say n_2, needs to satisfy the sample complexity requirement for the estimator to work, and the estimation error decreases as $\frac{c}{\sqrt{\min(n_1,n_2)}}$, where c depends on the Gaussian width of the unit norm ball. For example, for ℓ_1 norm applied to s-sparse change, the change can be accurately estimated with $\min(n_1, n_2) = O(s \log p)$ which is sharper than an existing result $n_1 = O(s^2 \log p)$ and $n_2 = O(n_1^2)$. Experimental results illustrating the effectiveness of the proposed estimator are presented.

1. Introduction

Over the past decade, considerable progress has been made on estimating the statistical dependency structure of graphical models based on samples drawn from the model. In particular, such advances have been made for Gaussian graphical models, Ising models, Gaussian copulas, as well as certain multi-variate extensions of general exponential family distributions including multivariate Poisson models (Banerjee et al., 2008; Kanamori et al., 2009; Meinshausen & Bühlmann, 2006; Ravikumar et al., 2010; 2011; Yang et al., 2012).

In this paper, we consider Ising models and focus on the problem of *estimating changes in Ising model structure*: given two sets of samples $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ and $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$ respectively drawn from two p-dimensional Ising models with true parameters θ_1^* and θ_2^*, where $\theta_1^*, \theta_2^* \in \mathbb{R}^{p \times p}$, the goal is to estimate the change $\delta\theta^* = (\theta_1^* - \theta_2^*)$. In particular, we focus on the situation when the change $\delta\theta^*$ has structure, such as sparsity, block sparsity, or node-perturbed sparsity, which can be characterized by a suitable (atomic) norm (Chandrasekaran et al., 2012; Mohan et al., 2014). However, the individual model parameters θ_1^*, θ_2^* need not have any specific structure, and they may both correspond to dense matrices. The goal is to get an estimate $\delta\hat{\theta}$ of the change $\delta\theta^*$ such that the estimation error $\Delta = (\delta\hat{\theta} - \delta\theta^*)$ is small. Such change estimation has potentially wide range of applications including identifying the changes in the neural connectivity networks, the difference between plant trait interactions at different climate conditions, and the changes in the stock market dependency structures.

One can consider two broad approaches for solving such change estimation problems: (i) *indirect change estimation*, where we estimate $\hat{\theta}_1$ and $\hat{\theta}_2$ from two sets of samples separately and obtain $\delta\hat{\theta} = (\hat{\theta}_1 - \hat{\theta}_2)$, or (ii) *direct change estimation*, where we directly estimate $\delta\hat{\theta}$ using the two sets of samples, without estimating θ_1 and θ_2 individually. In a high dimensional setting, recent advances (Cai et al., 2011; Ravikumar et al., 2010; 2011) illustrate that accurate estimation of the parameter θ^* of an Ising model depends on how sparse or otherwise structured the true parameter θ^* is. For example, if both θ_1^* and θ_2^* are sparse and the samples n_1, n_2 are sufficient to estimate them accurately (Ravikumar et al., 2010), indirect estimation of $\delta\hat{\theta}$ should be accurate. However, if the individual parameters θ_1^* and θ_2^* are somewhat dense, and the change $\delta\theta^*$ has considerably more structure, such as block sparsity (only a small block has changed) or node perturbation sparsity (only edges from a few nodes have changed) (Mohan et al., 2014), direct estimation may be considerably more efficient

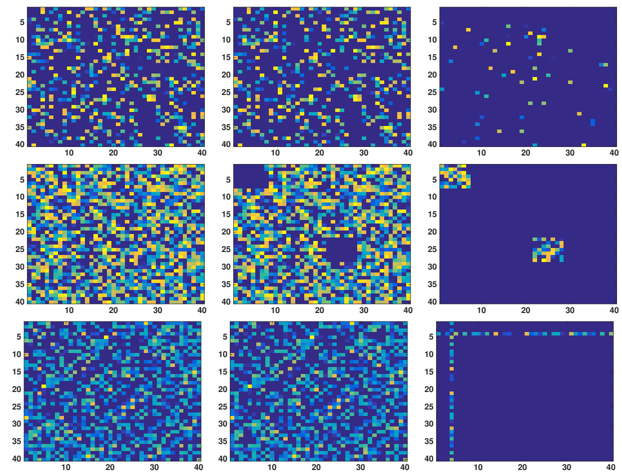
Proceedings of the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48.

both in terms of the number of samples required as well as the computation time.

Related Work: In recent work, Liu et al. (2015a) proposed a direct change estimator for graphical models based on the ratio of the probability density of the two models (Gretton et al., 2009; Kanamori et al., 2009; Sugiyama et al., 2008; 2012; Vapnik & Izmailov, 2015). They focused on the special case of L_1 norm, i.e., $\delta\theta^* \in \mathbb{R}^{p^2}$ is sparse, and provided non-asymptotic error bounds for the estimator along with a sample complexity of $n_1 = O(s^2 \log p)$ and $n_2 = O(n_1^2)$ for an unbounded density ratio model, where s is the number of the changed edges with p being the number of variables. Liu et al. (2015b) improved the sample complexity to $\min(n1, n2) = O(s^2 \log p)$ when a bounded density ratio model is assumed. Zhao et al. (2014) considered estimating direct sparse changes in Gaussian graphical models (GGMs). Their estimator is specific to GGMs and can not be applied to Ising models.

Our Contributions: We consider general structured direct change estimation, while allowing the change to have any structure which can be captured by a suitable (atomic) norm $R(\cdot)$. Our work is a considerable generalization of the existing literature which can only handle sparse changes, captured by the L_1 norm. In particular, our work now enables estimators for more general structures such as group/block sparsity, hierarchical group/block sparsity, node perturbation based sparsity, and so on (Banerjee et al., 2014; Chandrasekaran et al., 2012; Mohan et al., 2014; Negahban et al., 2012). Interestingly, for the unbounded density ratio model, our analysis yields sharper bounds for the special case of ℓ_1 norm, considered by Liu et al. (2015a). In particular, when $\delta\theta^*$ is sparse and our estimator is run with L_1 norm, we get a sample complexity of $n_1 = n_2 = O(s \log p)$ which is sharper than $n_1 = O(s^2 \log p)$ and $n_2 = O(n_1^2)$ in (Liu et al., 2015a).

The regularized estimator we analyze is broadly a Lasso-type estimator, with key important differences: the objective *does not* decompose additively over the samples, and the objective depends on samples from two distributions. The estimator builds on the density ratio estimator in (Liu et al., 2015a), but works with general norm regularization (Banerjee et al., 2014; Chandrasekaran et al., 2012; Negahban et al., 2012) where the regularization parameter λ_{n_1,n_2} depends on the sample size for both Ising models. Our analysis is quite different from the existing literature in change estimation. Liu et. (2015a) build on the primal-dual witness approach of Wainwright (Wainwright, 2009), which is effective for the special case of L_1 norm. Our analysis is largely geometric, where generic chaining (Talagrand, 2014) plays a key role, and our results are in terms of Gaussian widths of suitable sets associated with the norm (Banerjee et al., 2014; Chandrasekaran et al., 2012).



(g) θ_1 (h) θ_2 (i) $\delta\theta = \theta_1 - \theta_2$

Figure 1. Examples of $\delta\theta$ with different structures. First, second, and last rows show the sparsity structure of $\delta\theta$. Second row presents the group sparsity structure. Last row shows the node perturbation structure. Blue represents zeros.

The rest of the paper is organized as follows. In Section 2, we introduce the direct change estimator based on the ratio of the probability density of the Ising models. In Section 3, we establish statistical consistency of the direct change estimator, and conclude in Section 5.

2. Generalized Direct Change Estimation

We consider the following optimization problem

$$\underset{\delta\theta}{\operatorname{argmin}} \ \mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) + \lambda_{n_1,n_2} R(\delta\theta), \tag{1}$$

where $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ and $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$ are two sets of i.i.d binary samples drawn from from Ising graphical models with parameter θ_1^* and θ_2^*, respectively, each $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ are $p-$dimensional vectors, and n_1, n_2 are the respective sample sizes.

In this Section, we first give a brief background on Ising model selection. Then, we explain how to develop the loss function $\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})$ based on the density ratio (Gretton et al., 2009; Kanamori et al., 2009; Sugiyama et al., 2008; Vapnik & Izmailov, 2015) to directly estimate $\delta\theta = \theta_1 - \theta_2$, and finally we describe how to solve the optimization problem (1) for any norm $R(\delta\theta)$.

2.1. Ising Model

Let $X = (X_1, X_2, \cdots, X_p)$ denote a random vector in which each variable $X_s \in \{-1, 1\}$. Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \cdots, p\}$ and edge set E whose elements are unordered pairs of distinct vertices. The pairwise Ising Markov random field associated with the graph G over the random vector X is

$$P(X = \mathbf{x}|\theta^*) = \frac{1}{Z(\theta^*)} \exp\{ \sum_{s,t \in E} \theta^*_{s,t} x_s x_t \} \quad (2)$$

$$= \frac{1}{Z(\theta^*)} \exp\{ \langle \theta^*, T(\mathbf{x}) \rangle \} \quad (3)$$

$$= \frac{1}{Z(\Theta^*)} \exp\{ \mathbf{x}^T \Theta^* \mathbf{x} \} \quad (4)$$

where $T(\mathbf{x}) = \{x_s x_t\}_{s,t=1}^p$ is a vector of size $m = p^2$, $\theta^* = \{\theta^*_{s,t}\}_{s,t=1}^p \in \mathbb{R}^m$ and $\langle ., . \rangle$ is the inner product operator, and $\Theta^* \in \mathbb{R}^{p \times p}$ where $\Theta^*_{s,t} = \theta^*_{s,t}$. Note that basic Ising models also have non-interacting terms like $\alpha_s x_s$ and we are assuming these terms are zero, and they do not affect the dependency structure.

The parameter θ^* associated with the structure of the graph G reveals the statistical conditional independence structure among the variables i.e., if $\theta^*_{s,t} = 0$, then feature X_s is conditionally independent of X_t given all other variables and there is no edge in the graph G.

The partition function, $Z(\theta^*)$, plays the role of a normalizing constant, ensuring that the probabilities add up to one which is defined as

$$Z(\theta^*) = \sum_{\mathbf{x} \in \mathcal{X}} \exp\{ \langle \theta^*, T(\mathbf{x}) \rangle \} = \exp\{ \Psi(\theta^*) \}, \quad (5)$$

where $\mathcal{X}$ be the set of all possible configurations of X.

2.2. Loss Function $\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})$

Here, we build the loss function based on equation (3). Similarly, one can rewrite the loss function based on (4) if the regularization function is over matrices.

Consider two Ising models with parameters $\theta_1^* \in \mathbb{R}^{p^2}$ and $\theta_2^* \in \mathbb{R}^{p^2}$. Following Liu et. al (Liu et al., 2014; 2015a), a direct estimate for the changes detection problem based on density ratio can be posed as follows

$$r(X = \mathbf{x}|\delta\theta) = \frac{p(X = \mathbf{x}|\theta_1)}{p(X = \mathbf{x}|\theta_2)} = \underbrace{\frac{\exp\{\langle T(\mathbf{x}), \theta_1 \rangle\}}{\exp\{\langle T(\mathbf{x}), \theta_2 \rangle\}}}_{r^*(\mathbf{x}|\delta\theta)} \underbrace{\frac{Z(\theta_2)}{Z(\theta_1)}}_{1/Z(\delta\theta)}$$

$$= \frac{\exp\{\langle T(\mathbf{x}), \delta\theta \rangle)\}}{Z(\delta\theta)}, \quad (6)$$

where the parameter $\delta\theta = \theta_1 - \theta_2$ encodes the change between two graphical models θ_1 and θ_2.

First, we show that $Z(\delta\theta) = E_{X \sim p(X|\theta_2)}[e^{\langle T(X), \delta\theta \rangle}]$:

$$Z(\delta\theta) = \frac{Z(\theta_1)}{Z(\theta_2)} = \frac{1}{Z(\theta_2)} \sum_{\mathbf{x} \in \mathcal{X}} e^{\langle T(\mathbf{x}), \theta_1 \rangle} \quad (7)$$

$$= \frac{1}{Z(\theta_2)} \sum_{\mathbf{x} \in \mathcal{X}} e^{\langle T(\mathbf{x}), \theta_2 \rangle} \frac{e^{\langle T(\mathbf{x}), \theta_1 \rangle}}{e^{\langle T(\mathbf{x}), \theta_2 \rangle}}$$

$$= \sum_{\mathbf{x} \in \mathcal{X}} \underbrace{\frac{e^{\langle T(\mathbf{x}), \theta_2 \rangle}}{Z(\theta_2)}}_{p(\mathbf{x}|\theta_2)} e^{\langle T(\mathbf{x}), \delta\theta \rangle} = E_{X \sim p(X|\theta_2)}[e^{\langle T(X), \delta\theta \rangle}].$$

Next, using the samples $\mathfrak{X}_2^{n_2}$ from $p(X|\theta_2)$, we estimate $Z(\delta\theta)$ empirically as

$$\hat{Z}(\delta\theta) = \frac{1}{n_2} \sum_{i=1}^{n_2} \exp\{ \langle T(\mathbf{x}_i^2), \delta\theta \rangle \}, \quad (8)$$

and the sample approximation of $r(X|\delta\theta)$ is given as

$$\hat{r}(X = \mathbf{x}|\delta\theta) = \frac{r^*(X = \mathbf{x}|\theta_1)}{\hat{Z}(\delta\theta)}$$

$$= \frac{\exp\{\langle T(\mathbf{x}), \delta\theta \rangle\}}{\frac{1}{n_2} \sum_{i=1}^{n_2} \exp\{ \langle T(\mathbf{x}_i^2), \delta\theta \rangle \}}. \quad (9)$$

Using the fact that $r(X|\delta\theta^*) q(X|\theta_2^*) = p(X|\theta_1^*)$, we approximate $\hat{r}(X|\delta\theta)$, by minimizing the KL divergence,

$$KL\left(p(X|\theta_1^*) \| \hat{r}(X|\delta\theta) p(X|\theta_2^*) \right) \quad (10)$$

$$= \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}|\theta_1^*) \log \frac{p(\mathbf{x}|\theta_1^*)}{p(\mathbf{x}|\theta_2^*) \hat{r}(\mathbf{x}|\delta\theta)}$$

$$= \underbrace{KL\left(p(X|\theta_1^*) \| p(X|\theta_2^*) \right)}_{\text{Constant}} - E_{X \sim p(X|\theta_1^*)} \left[\log \hat{r}(X|\delta\theta) \right]$$

Thus, using the samples $\mathfrak{X}_1^{n_1}$ and $\mathfrak{X}_2^{n_2}$, we define the empirical loss function

$$\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) = \frac{-1}{n_1} \sum_{i=1}^{n_1} \log \hat{r}(\mathbf{x}_i^1|\delta\theta) \quad (11)$$

$$= \frac{-1}{n_1} \sum_{i=1}^{n_1} \langle T(\mathbf{x}_i^1), \delta\theta \rangle + \underbrace{\log \frac{1}{n_2} \sum_{i=1}^{n_2} \exp\{ \langle T(\mathbf{x}_i^2), \delta\theta \rangle \}}_{\hat{\Psi}(\delta\theta)}$$

Remark 1 Note that the loss function (11) does not additively decompose over the samples. The second term in (11) is the logarithm over sum of a function of samples.

2.3. Optimization

The optimization problem (1) has a composite objective with a smooth convex term corresponding to the loss function (11) and a a potentially non-smooth convex term corresponding to the regularizer. In this section, we present an algorithm in the class of Fast Iterative Shrinkage-Thresholding Algorithms (FISTA) for efficiently solving the problem (1) (Beck & Teboulle, 2009). For convenience, we refer the loss function $\mathcal{L}(\delta\theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})$ as $\mathcal{L}(\delta\theta)$ and we drop the subscript $\{n_1, n_2\}$ of λ_{n_1,n_2}.

One of the most popular methods for composite objective functions is in the class of FISTA where at each iteration we linearize the smooth term and minimize the quadratic approximation of the form

$$\begin{aligned} Q_L(\delta\theta, \delta\theta_t) :=& \mathcal{L}(\delta\theta) + \langle \delta\theta - \delta\theta_t, \nabla\mathcal{L}(\delta\theta_t)\rangle \\ &+ \frac{L}{2}\|\delta\theta - \delta\theta_t\|_2^2 + \lambda R(\delta\theta), \end{aligned} \tag{12}$$

where L denotes the Lipschitz constant of the loss function $\mathcal{L}(\delta\theta)$. Ignoring constant terms in $\delta\theta_t$, the unique minimizer of the above expression (12) can be written as

$$\begin{aligned} p_L(\delta\theta_t) &= \arg\min_{\delta\theta} Q_{\mathcal{L}}(\delta\theta, \delta\theta_t) \\ &= \arg\min_{\delta\theta} \lambda R(\delta\theta) + \frac{L}{2}\left\|\delta\theta - \left(\delta\theta_t - \frac{1}{L}\nabla\mathcal{L}(\delta\theta_t)\right)\right\|_2^2 \\ &= \arg\min_{\delta\theta} \frac{\lambda}{L} R(\delta\theta) + \frac{1}{2}\left\|\delta\theta - \left(\delta\theta_t - \frac{1}{L}\nabla\mathcal{L}(\delta\theta_t)\right)\right\|_2^2. \end{aligned} \tag{13}$$

In fact, the updates of $\delta\theta$ is to compute certain proximal operators of the non-smooth term $R(.)$. In general, the proximal operator $\text{prox}_h(\mathbf{x})$ of a closed proper convex function $h : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ is defined as

$$\text{prox}_h(\mathbf{x}) = \underset{\mathbf{u}}{\text{argmin}} \left(h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2\right). \tag{14}$$

Thus, the unique minimizer (13) correspond to $\text{prox}_{\frac{\lambda}{L}R}\left(\delta\theta_t - \frac{1}{L}\nabla\mathcal{L}(\delta\theta_t)\right)$ which has rate of convergence of $O(1/t)$ (Nesterov, 2005; Parikh & Boyd, 2014).

To improve the rate of convergence, we adapt the idea of FISTA algorithm. The main idea is to iteratively consider the proximal operator prox(.) at a specific linear combination of the previous two iterates $\{\delta\theta_t, \delta\theta_{t-1}\}$

$$\xi_{t+1} = \delta\theta_t + \alpha_{t+1}\left(\delta\theta_t - \delta\theta_{t-1}\right), \tag{15}$$

instead of just the previous iterate $\delta\theta_t$. The choice of α_{t+1} follows Nesterovs accelerated gradient descent (Nesterov, 2005; Parikh & Boyd, 2014) and is detailed in Algorithm 1. The iterative algorithm simply updates

$$\delta\theta_{t+1} = \text{prox}_{\frac{\lambda}{L}R}\left(\xi_{t+1} - \frac{1}{L}\nabla\mathcal{L}(\xi_{t+1})\right). \tag{16}$$

The algorithm has a rate of convergence of $O(1/t^2)$ (Beck & Teboulle, 2009).

Algorithm 1 Generalized Direct Change Estimator

Input: $L_0 > 0$, $\mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}$
Step 0. Set $\xi_1 = \delta\theta_0$, $t = 1$
Step t. ($t \geq 1$) Find the smallest non-negative integers i_t such that with $\tilde{L} = 2^{i_t}L_{t-1}$

$$\mathcal{L}\left(p_{\tilde{L}}(\xi_t)\right) + R\left(p_{\tilde{L}}(\xi_t)\right) \leq Q_{\tilde{L}}\left(p_{\tilde{L}}(\xi_t), \xi_t\right). \tag{17}$$

Set $L_t = 2^{i_t}L_{t-1}$ and Compute

$$\delta\theta_t = \text{prox}_{\frac{\lambda}{L}R}\left(\xi_t - \frac{1}{L}\nabla\mathcal{L}(\xi_t)\right) \tag{18}$$

$$\beta_{t+1} = \frac{1 + \sqrt{1 + 4\beta_t^2}}{2} \tag{19}$$

$$\xi_{t+1} = \delta\theta_t + \left(\frac{\beta_t - 1}{\beta_{t+1}}\right)(\delta\theta_t - \delta\theta_{t-1}) \tag{20}$$

2.4. Regularization Function $\mathbf{R}(\boldsymbol{\delta\theta})$

We assume that the optimal $\delta\theta^*$ is sparse or suitably 'structured' where such structure can be characterized by having a low value according to a suitable norm $R(\delta\theta^*)$. In below, we provide a few examples of such a norm.

L_1 norm: One example for $R(.)$ we will consider throughout the paper is the L_1 norm regularization. We use L_1 norm if only a few edges has changed (1st row in Figure 1). In particular, we consider $R(\delta\theta) = \|\delta\theta\|_1$ if number of non-zeros entries in $\delta\theta^*$ is $s < p^2$. The $\text{prox}_{\frac{\lambda}{L}\|\cdot\|_1}(.)$ is given by the elementwise soft-thresholding operation (Singer & Duchi, 2009) as

$$\left[\text{prox}_{\frac{\lambda}{L}\|\cdot\|_1}\right]_i(\mathbf{z}) = \text{sign}(\mathbf{z}_i).\max(0, \mathbf{z}_i - \frac{\lambda}{L}). \tag{21}$$

Group-sparse norm: Another popular example we consider is the group-sparse norm. We use group lasso norm if a group of edges has changed (2nd row in Figure 1). For some kinds of data, it is reasonable to assume that the variables can be clustered (or grouped) into types, which share similar connectivity or correlation patterns. Let $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_{N_G}\}$ denote a collection of groups, which are subsets of variables. We assume that $\delta\Theta^*(s,t) = 0$ for any variable $s \in G_g$ and for any variable $t \in G_h$. In the group sparse setting for any subset $S_{\mathcal{G}} \subseteq \{1, 2, \cdots N_G\}$ with cardinality $|S_{\mathcal{G}}| = s_{\mathcal{G}}$, we assume that the parameter $\delta\Theta^*$ satisfies $\{\delta\Theta^*_{s,t} = 0 : s,t \in G_g \;\&\; g \notin S_{\mathcal{G}}\}$. We will focus on the case when $R(\delta\Theta) = \sum_{g=1}^{N_G} \|\delta\Theta(s,t) : s,t \in G_g\|_F$ (Marlin et al., 2009). Let $\delta\Theta_{G_g}$ bd the sub-matrix of $\delta\Theta$ covering nodes in G_g. Proximal operator is given by the group specific soft-thresholding operation.

$$\left[\text{prox}_{\frac{\lambda}{L}R}\right]_g(\delta\Theta) = \frac{\max(\|\delta\Theta_{G_g}\|_F - \frac{\lambda}{L}, 0)}{\|\delta\Theta_{G_g}\|_F}. \tag{22}$$

Node perturbation: Another example is the row-column overlap norm (RCON) (Mohan et al., 2014) to capture perturbed nodes i.e., nodes that have a completely different connectivity pattern to other nodes among two networks (3rd row in Figure 1). A special case of RCON we are interested is $\sum_{i=1}^{p} \|V_i\|_q$ where $\delta\Theta = V + V^T$, and V_i is the $i-$th column of matrix V. This norm can be viewed as overlapping group lasso (Mohan et al., 2014) and thus can be solved by applying Algorithm 1 with proximal operator for overlapping group lasso (Yuan et al., 2011). Also, we can write problem (1) as a constrained optimization

$$\begin{aligned} &\underset{\delta\Theta, V}{\operatorname{argmin}} \;\; \mathcal{L}(\delta\Theta; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}) + \lambda_1 \delta\Theta\|_1 + \lambda_{n_1,n_2} \sum_{i=1}^{p} \|V_i\|_q \\ &\text{s.t} \qquad \delta\Theta = V + V^T, \end{aligned} \tag{23}$$

and solve it by applying in-exact ADMM techniques (Mohan et al., 2014).

3. Statistical Recovery for Generalized Direct Change Estimation

Our goal is to provide non-asymptotic bounds on $\|\Delta\|_2 = \|\delta\theta^* - \delta\hat{\theta}\|_2$ between the true parameter $\delta\theta^*$ and the minimizer $\delta\hat{\theta}$ of (1). In this section, we describe various aspects of the problem, introducing notations along the way, and highlight our main result.

3.1. Background and Assumption

Gaussian Width: In several of our proofs, we use the concept of Gaussian width (Chandrasekaran et al., 2012; Gordon, 1988), which is defined as follows.

Definition 1 For any set $A \in \mathbb{R}^p$, the *Gaussian width* of the set A is defined as:

$$w(A) = E_g \left[\sup_{u \in A} \langle g, u \rangle \right] . \tag{24}$$

where the expectation is over $g \sim N(0, \mathbb{I}_{p \times p})$, a vector of independent zero-mean unit-variance Gaussian random variable.

The Gaussian width $w(A)$ provides a geometric characterization of the size of the set A. Consider the Gaussian process $\{Z_u\}$ where the constituent Gaussian random variables $Z_u = \langle u, g \rangle$ are indexed by $u \in A$, and $g \sim N(0, \mathbb{I}_{p \times p})$. Then the Gaussian width $w(A)$ can be viewed as the expectation of the supremum of the Gaussian process $\{Z_u\}$. Bounds on the expectations of Gaussian and other empirical processes have been widely studied in the literature, and we will make use of generic chaining for some of our analysis (Boucheron et al., 2013; Ledoux & Talagrand, 2013; Talagrand, 2005; 2014).

The Error Set: Consider solving the problem (1), under assumption $\lambda_{n_1,n_2} > \beta R^* \left(\nabla \mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)$, where $\beta > 1$ and $R^*(.)$ is the dual norm of $R(.)$. Banerjee et al. (Banerjee et al., 2014) show that for any convex loss function the error vector $\Delta = (\delta\theta^* - \delta\hat{\theta})$ lies in a restricted set that is characterized as

$$\begin{aligned} E_r &= E_r(\delta\theta^*, \beta) \\ &= \left\{ \Delta \in \mathbb{R}^p \;\middle|\; R(\delta\theta^* + \Delta) \le R(\delta\theta^*) + \frac{1}{\beta} R(\Delta) \right\} . \end{aligned} \tag{25}$$

Restricted Strong Convexity (RSC) Condition: The sample complexity of the problem (1) depends on the RSC condition (Negahban et al., 2012), which ensures that the estimation problem is strongly convex in the neighborhood of the optimal parameter (Banerjee et al., 2014; Negahban et al., 2012). A convex loss function satisfies the RSC condition in $C_r = cone(E_r)$, i.e., $\forall \Delta \in C_r$, if there exists a suitable constant κ such that

$$\begin{aligned} \delta\mathcal{L}(\delta\theta^*, u) :=& \mathcal{L}(\delta\theta^* + u) - \mathcal{L}(\delta\theta^*) - \langle \nabla \mathcal{L}(\delta\theta^*), u \rangle \\ \ge& \kappa \|u\|_2^2 \end{aligned} \tag{26}$$

Deterministic Recovery Bounds: If the RSC condition is satisfied on the error set C_r and λ_{n_1,n_2} satisfies the assumptions stated earlier, for any norm $R(.)$, Banerjee et al. (Banerjee et al., 2014) show a deterministic upper bound for $\|\Delta\|_2$ in terms of λ_{n_1,n_2}, κ, and the norm compatibility constant $\Psi(C_r) = \sup_{\mathbf{u} \in C_r} \frac{R(\mathbf{u})}{\|\mathbf{u}\|_2}$, as

$$\|\Delta\|_2 \le \frac{1+\beta}{\beta} \frac{\lambda_{n_1,n_2}}{\kappa} \Psi(C_r) \, . \tag{27}$$

Smooth Density Ratio Model Assumption: For any vector $\mathbf{u}$ such that $\|\mathbf{u}\|_2 \le \|\delta\theta^*\|_2$ and every $\epsilon \in R$, the following inequality holds:

$$E_{X \sim p(X|\theta_2)}[\exp\{\epsilon \, r(X|\delta\theta^* + \mathbf{u}) - 1\}] \le \exp\{\epsilon^2\}.$$

A similar assumption is used in the analysis of Liu et al. (Liu et al., 2015a).

Remark 2 Bounded density ratio is a special case satisfying the smooth density ratio assumption. Lemma 1 shows a sufficient condition under which the density ratio is bounded.

Lemma 1 *Consider two Ising Model with true parameters θ_1^* and θ_2^*. Let $d_1, d_2 \gg s$ where $\|\theta_1^*\|_0 = d_1$, $\|\theta_2^*\|_0 = d_2$, and $\|\delta\theta^*\|_0 = s$. Assume*

$$\min_{i,j=1\cdots p} (|\theta_1^*(i,j)|) \ge \frac{1}{d_1 - 1} - \frac{c_1}{(d_1-1)s} \tag{28}$$

$$\min_{i,j=1\cdots p} (|\theta_2^*(i,j)|) \ge \frac{1}{d_2 - 1} - \frac{c_2}{(d_2-1)s}, \tag{29}$$

where c_1 and c_2 are positive constants. Then the density ratio $r(X = \mathbf{x}|\delta\theta^*)$ is bounded.

Note that if individual graphs are dense, then the conditions (28) and (29) are satisfied and as a result the smooth density ratio is satisfied.

Remark 3 In this paper, we focus on the Ising graphical model. But, our statistical analysis holds for any graphical models that satisfy the above mentioned assumption. Through our analysis, no assumption is required on the individual graphical models.

3.2. Bounds on the regularization parameter

To get the recovery bound (27) above, one needs to have $\lambda_{n_1,n_2} \geq \beta R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)$. However, the bound on λ_{n_1,n_2} depends on unknown quantity $\delta\theta^*$ and the samples $\mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}$ and is hence random. To overcome the above challenges, one can bound the expectation $E[R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)]$ over all samples of size n_1 and n_2, and obtain high-probability deviation bounds. The goal is to provide a sharp bound on λ_{n_1,n_2} since the error bound in (27) is directly proportional to λ_{n_1,n_2}.

In theorem 1, we characterize the expectation $E[R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right)]$ in terms of the Gaussian width of the unit norm-ball of $R(.)$, which leads to a sharp bound. The upper bound on Gaussian width of the unit norm-ball of R for atomic norms which covers a wide range of norms is provided in (Chandrasekaran et al., 2012; Chen & Banerjee, 2015).

Theorem 1 *Define $\Omega_R = \{u : R(u) \leq 1\}$. Let $\phi(R) = \sup_{\mathbf{u}} \frac{\|\mathbf{u}\|_2}{R(\mathbf{u})}$. Assume that for any $\mathbf{u}$ that $\|\mathbf{u}\| \leq \|\theta^*\|$*

$$\frac{1}{2}\lambda_{\max}\left(\nabla^2\mathcal{L}(\delta\theta^* + \mathbf{u})\right) \leq \eta_0, \tag{30}$$

where $\lambda_{\max}(.)$ is the maximum eigenvalue. Then under the smooth density ratio assumption, we have

$$E\left[R^*(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2}))\right] \leq \frac{2\sqrt{\eta_0}(c_1 w\left(\Omega_R\right) + \phi(R))}{\sqrt{\min(n_1, n_2)}}.$$

and with probability at least $1 - c_2 e^{-\epsilon^2}$

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right) \leq \frac{c_2(1+\epsilon)w(\Omega_R) + \tau_1}{\sqrt{\min(n_1, n_2)}}.$$

where c_1 and c_2 are positive constants, $\tau_1 = 2\sqrt{\eta_0}\phi(R)$, and $w(\Omega_R)$ is the Gaussian width of set Ω_R.

Note, that our analysis hold for any norm and it is expressed in terms of the Gaussian width. In the following, we give the bound on the regularization parameter for two examples of the regularization function $R(.)$.

Corollary 1 *If $R(\delta\theta)$ is the L_1 norm, and $\delta\theta \in \mathbb{R}^{p^2}$ then with high probability we have the bound*

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right) \leq \frac{\eta_2\sqrt{\log p}}{\sqrt{\min(n_1, n_2)}}. \tag{31}$$

Corollary 2 *If $R(\delta\theta)$ is the group-sparse norm, and $\delta\theta \in \mathbb{R}^{p^2}$ then with high probability we have the bound*

$$R^*\left(\nabla\mathcal{L}(\delta\theta^*; \mathfrak{X}_1^{n_1}, \mathfrak{X}_2^{n_2})\right) \leq \frac{\eta_2\sqrt{m + \log N_G}}{\sqrt{\min(n_1, n_2)}}, \tag{32}$$

where $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_{N_G}\}$ is a collection of groups, $m = \max_i |\mathcal{G}_i|$ is the maximum size of any group.

3.3. RSC Condition

In this Section, we establish the RSC condition for direct change detection estimator (1). Simplifying the expression and applying mean value theorem twice on the left side of RSC condition (26), for $\forall\gamma_i \in [0, 1]$, we have

$$\begin{aligned}\delta\mathcal{L}(\delta\theta^*, u) &:= \mathcal{L}(\delta\theta^* + u) - \mathcal{L}(\delta\theta^*) - \langle\nabla\mathcal{L}(\delta\theta^*), u\rangle \\ &\geq u^T\nabla^2\mathcal{L}(\delta\theta^* + \gamma_i u)u.\end{aligned} \tag{33}$$

Thus, the RSC condition depends on the non-linear terms of loss function. Recall that the nonlinear term, second term, in Loss function (1) which is the approximation of the log-partition functions only depends on n_2. As a results, only samples of $\mathfrak{X}_2^{n_2}$ affect the RSC conditions. Our analysis is an extension of the results on (Banerjee et al., 2014) using the generic chaining. We show that, with high probability the RSC condition is satisfied once samples n_2 crosses $w^2(C_r \cap S^{d-1})$ the Gaussian width of restricted error set. The bound on Gaussian width of the error set for atomic norms has been provided in (Chen & Banerjee, 2015).

Let $r_i = r(X = \mathbf{x_i^2}|\delta\theta^*)$ and $\bar{\varepsilon}$ denote the probability that r_i exceeds some constant T: $\bar{\varepsilon} = p(r_i > T) \leq 2e^{-\frac{T^2}{2}}$.

Theorem 2 *Let $X \in \mathbb{R}^{n\times p}$ be a design matrix with independent isotropic sub-Gaussian rows with $|||X_i|||_{\Psi_2} \leq \kappa$. Then, for any set $A \subseteq S^{p-1}$, for suitable constants η, c_1, $c_2 > 0$ with probability at least $1 - \exp(-\eta w^2(A))$, we have*

$$\inf_{u\in A}\partial\mathcal{L}(\theta^*; u, X) \geq c_1\underline{\rho}^2\left(1 - c_2\kappa_1^2\frac{w(A)}{\sqrt{n_2}}\right) - \tau \tag{34}$$

where $\kappa_1 = \frac{\kappa}{\bar{\varepsilon}}$, $\underline{\rho}^2 = \inf_{\mathbf{u}\in A}\rho_{\mathbf{u}}^2$ with $\rho_{\mathbf{u}}^2 = E\left[\langle\mathbf{u}, T(X_i^2)\rangle^2\mathbb{I}(r_i > T)\right]$, and τ is smaller than the first term in right hand side. Thus, for $n_2 \geq c_2 w^2(A)$, with probability at least $1 - \exp(-\eta w^2(A))$, we have $\inf_{u\in A}\partial\mathcal{L}(\theta^*; u, X) > 0$.

3.4. Statistical Recovery

With the above results in place, from (27), Theorem 3 provides the main recovery bound for generalized direct change estimator (1).

Theorem 3 *Consider two set of i.i.d samples* $\mathfrak{X}_1^{n_1} = \{\mathbf{x}_i^1\}_{i=1}^{n_1}$ *and* $\mathfrak{X}_2^{n_2} = \{\mathbf{x}_i^2\}_{i=1}^{n_2}$. *Define* $\Omega_R = \{u : R(u) \leq 1\}$. *Assume that* $\delta\hat{\theta}$ *is the minimizer of the problem* (1). *Then, with probability at least* $1 - \eta_0 e^{-\epsilon^2}$ *the followings hold*

$$\lambda_{n_1,n_2} \geq \frac{\eta_1}{\sqrt{\min(n_1, n_2)}}(w(\Omega_R) + \epsilon) \qquad (35)$$

and for $n_2 \geq cw^2(C_r \cap S^{d-1})$, *with high probability, the estimate* $\delta\hat{\theta}$ *satisfies*

$$\|\Delta\|_2 \leq O\left(\frac{w(\Omega_R)}{\sqrt{\min(n_1, n_2)}}\right)\Psi(C_r)\ , \qquad (36)$$

where $w(.)$ *is the Gaussian width of a set,and* c_2, η_0, *and* η_1 *are positive constants.*

Proof: Proof of the Theorem can be directly obtain as the results of (27) and Theorem 1 and Theorem 2.

In the following, we provide the recovery bound for two special cases as an example.

Corollary 3 *If* $R(\delta\theta)$ *is the* L_1 *norm,* $\delta\theta^* \in \mathbb{R}^{p^2}$ *s s-sparse.,* $\Psi(C_r) \leq 4\sqrt{s}$, *and for* $n_2 > cs \log p$, *the recovery error is bounded by*

$$\|\Delta\|_2 \leq c_3 \frac{\Psi(C_r)\lambda_{n_1,n_2}}{\kappa} = O\left(\sqrt{\frac{s \log p}{\min(n_1, n_2)}}\right).$$

Corollary 4 *If* $R(\delta\theta)$ *is the group-sparse norm,* $\delta\theta \in \mathbb{R}^{p^2}$, $\Psi(C_r) \leq 4\sqrt{s_G}$ *and for* $n_2 \geq c(ms_G + s_G \log N_G)$, *the recovery error is bounded by*

$$\|\Delta\|_2 \leq c_3 \frac{\Psi(C_r)\lambda_{n_1,n_2}}{\kappa} = O\left(\sqrt{\frac{s_G m + \log N_G}{\min(n_1, n_2)}}\right).$$

4. Experiments

In this Section, we evaluate generalized direct change estimator (direct) with three different norms. and we compare our direct approach with indirect approach. For indirect approach, we first estimate Ising model structures $\hat{\theta}_1$ and $\hat{\theta}_2$ with L_1 norm regularizer, separately (Ravikumar et al., 2010). Then, we obtain $\delta\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$. In all experiments, we draw n_1 and n_2 i.i.d samples from each Ising model by running Gibbs sampling. Here we set $n = n_1 = n_2 = \{20, 50, 100\}$.

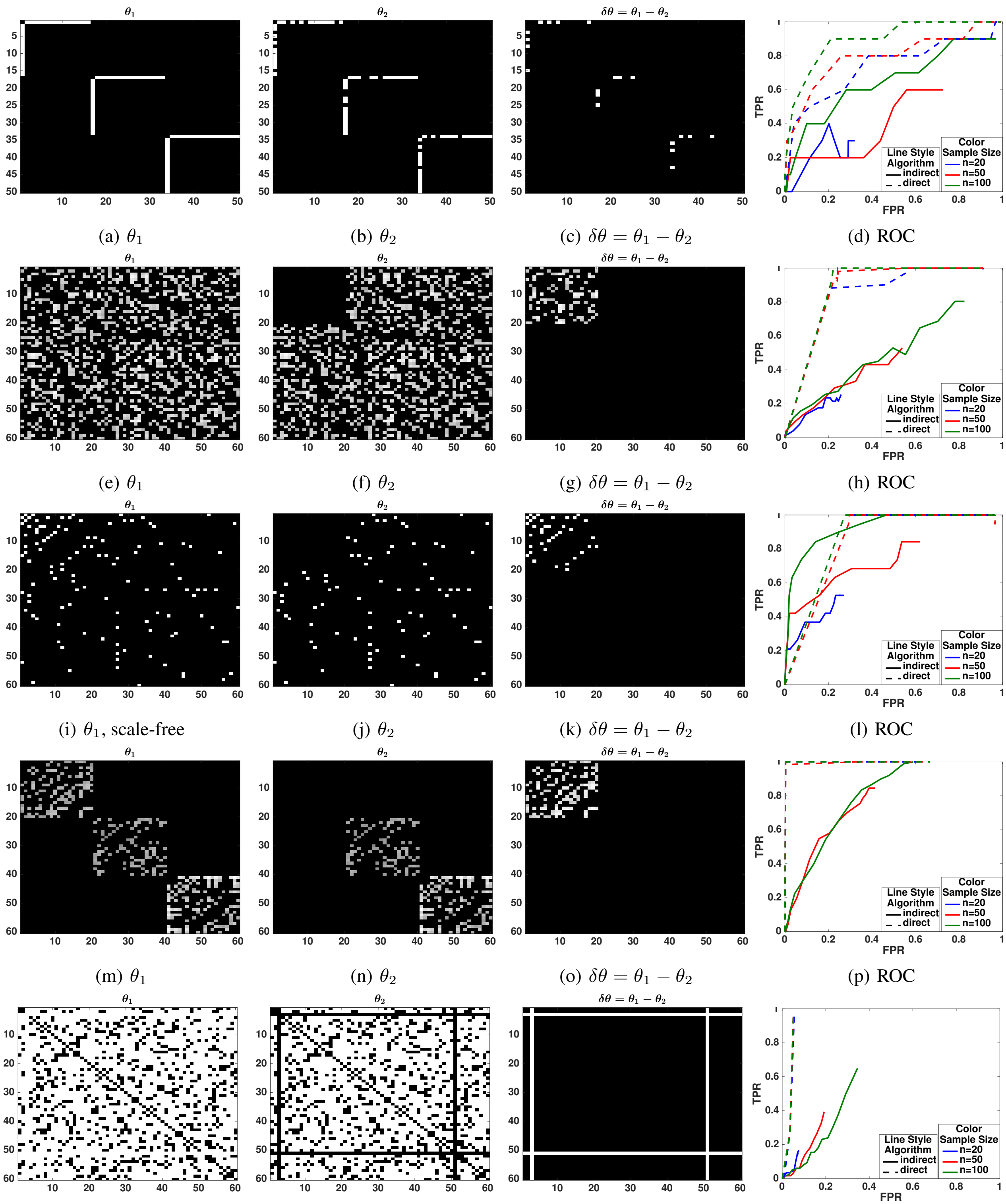
L_1 **norm:** Here we first generate θ_1^* with three disconnected star sub-graphs (Figure 4-a) with $p = 50$. We generate the weights uniformly random between $\{0.3 - 0.5\}$. We then generate θ_2^* by removing 10 random edges from θ_1^* (Figure 4-b). It is interesting that although individual graphs are sparse, but direct approach has a better ROC curve for all values of n (Figure 4-d). Similar results obtained by with random graph structure of θ_1^* and θ_2^*.

Group-sparse norm: In this set of experiments, we evaluate direct method with three different structure for θ_1^*: (i) a random graph structure (Figure 4-e), (ii) scale free graph structure (Figure 4-i), and (iii) a cluster graph structure (Figure 4-m). In all settings, we set $p = 60$ and generate θ_2^* by removing a block of edges from θ_1^* (Figure 4-(f,j,n)). For random graph structure and block structure, direct method has a better ROC curve (Figure 4-h,p). But, for scale-free structure, since the individual graphs are sparse, indirect method can estimate $\hat{\theta}_1$ and $\hat{\theta}_2$ correctly, and thus have a better ROC curve (Figure 4-l).

Node perturbation: Here, we first generate a random graph structure θ_1^*, and then generate θ_2^* by perturbing two nodes in θ_1^*. Here we set $p = 60$ and generate θ_2^* by setting rows and columns $3, 51$ to zero in θ_1^* (Figure 4-s). Although, the individual graphs are dense but direct approach can estimate edges in $\delta\theta$ with only $n = 20$ samples (Figure 4-t).

5. Conclusion

This paper presents the statistical analysis of direct change problem in Ising graphical models where any norm can be plugged in for characterizing the parameter structure. An optimization algorithm based on FISTA-style algorithms is proposed with the convergence rate of $O(1/T^2)$. We provide the statistical analysis for any norm such as L_1 norm, group sparse norm, node perturbation, etc. Our analysis is based on generic chaining and illustrates the important role of Gaussian widths (a geometric measure of size of suitable sets) in such results. For the special case of sparsity, we obtain a sharper result than previous results (Liu et al., 2015a) under the same smooth density ratio assumption. Liu et al. (Liu et al., 2015a) obtained the same result with a bounded density ratio assumption which is a more restrictive assumption. Although, we presented the results for Ising model, our analysis can be applied to any graphical model which satisfies the smooth density ratio assumption. Further, we extensively compared our generalized direct change estimator with an indirect approach over a wide range of graph structures and norms. We show that our direct approach has a better ROC curve than indirect approach without any assumption on the structure of individual graphs. We implemented indirect approach by estimating individual Ising model structures with L_1 norm regularizer. However, if individual graphs has a suitable structure such as group sparsity, one may apply a regularization that can characterize the graph structure and may improve performance of the indirect approach. We will investigate this possibility in our future research.



(a) θ_1 (b) θ_2 (c) $\delta\theta = \theta_1 - \theta_2$ (d) ROC

(e) θ_1 (f) θ_2 (g) $\delta\theta = \theta_1 - \theta_2$ (h) ROC

(i) θ_1, scale-free (j) θ_2 (k) $\delta\theta = \theta_1 - \theta_2$ (l) ROC

(m) θ_1 (n) θ_2 (o) $\delta\theta = \theta_1 - \theta_2$ (p) ROC

(q) θ_1 (r) θ_2 (s) $\delta\theta = \theta_1 - \theta_2$ (t) ROC

Figure 2. First row $\delta\theta^*$ has a sparse structure (L_1 norm) and θ_1^* has 3 disconnected star graphs. Second, third, and forth rows $\delta\theta^*$ has group sparse structure (group sparse norm) where θ_1^* has a random graph structure in second row, scale-free structure in third row, and block structure in forth row. Last row $\delta\theta^*$ has two perturbed norm (Node perturbation) and θ_1^* has a random graph structure. Blacks in heatmaps denotes zeros. ROC curve for different structures show in the last column. Direct approach has a better ROC curve for all structures except with scale-free structure of θ_1^*.

Acknowledgment

The research was supported by NSF grants IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS- 1314560, IIS-0953274, IIS-1029711, NASA grant NNX12AQ39A, and gifts from Adobe, IBM, and Yahoo. F. F. acknowledges the support of DDF (2015-2016) from the University of Minnesota.

References

Banerjee, Arindam, Chen, Sheng, Fazayeli, Farideh, and Sivakumar, Vidyashankar. Estimation with Norm Regularization. In *Neural Information Processing Systems*, 2014.

Banerjee, Onureena, El Ghaoui, Laurent, and d'Aspremont, Alexandre. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Cai, Tony, Liu, Weidong, and Luo, Xi. A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A., and Willsky, Alan S. The Convex Geometry of Linear Inverse Problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chen, Sheng and Banerjee, Arindam. Structured estimation with atomic norms: General bounds and applications. In *Neural Information Processing Systems*, 2015.

Gordon, Yehoram. On Milman's Inequality and Random Subspaces Which Escape Through a Mesh in Rn. In *Geometric Aspects of Functional Analysis*, volume 1317 of *Lecture Notes in Mathematics*, pp. 84–106. Springer Berlin, 1988.

Gretton, Arthur, Smola, Alex, Huang, Jiayuan, Schmittfull, Marcel, Borgwardt, Karsten, and Schölkopf, Bernhard. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Kanamori, Takafumi, Hido, Shohei, and Sugiyama, Masashi. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2013.

Liu, Song, Quinn, John A, Gutmann, Michael U, Suzuki, Taiji, and Sugiyama, Masashi. Direct learning of sparse changes in markov networks by density ratio estimation. *Neural computation*, 26(6):1169–1197, 2014.

Liu, Song, Suzuki, Taiji, and Sugiyama, Masashi. Support consistency of direct sparse-change learning in markov networks. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015a.

Liu, Song, Suzuki, Taiji, and Sugiyama, Masashi. Support consistency of direct sparse-change learning in markov networks. In *arXiv:1407.0581v10*, 2015b.

Marlin, Benjamin M, Schmidt, Mark, and Murphy, Kevin P. Group sparse priors for covariance estimation. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 383–392. AUAI Press, 2009.

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

Mohan, Karthik, London, Palma, Fazel, Maryam, Witten, Daniela, and Lee, Su-In. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

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

Nesterov, Yu. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Parikh, Neal and Boyd, Stephen P. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Ravikumar, Pradeep, Wainwright, Martin J, and Lafferty, John D. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Ravikumar, Pradeep, Wainwright, Martin J, Raskutti, Garvesh, and Yu, Bin. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Singer, Yoram and Duchi, John C. Efficient learning using forward-backward splitting. In *Neural Information Processing Systems*, pp. 495–503, 2009.

Sugiyama, Masashi, Nakajima, Shinichi, Kashima, Hisashi, Buenau, Paul V, and Kawanabe, Motoaki. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pp. 1433–1440, 2008.

Sugiyama, Masashi, Suzuki, Taiji, and Kanamori, Takafumi. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Talagrand, M. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

Talagrand, Michel. *The Generic Chaining*. Springer Monographs in Mathematics. Springer Berlin, 2005.

Vapnik, Vladimir and Izmailov, Rauf. Statistical inference problems and their rigorous solutions. In *Statistical Learning and Data Sciences*, pp. 33–71. Springer International Publishing, 2015.

Wainwright, Martin J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programmming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2201, 2009.

Yang, Eunho, Genevera, Allen, Liu, Zhandong, and Ravikumar, Pradeep K. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pp. 1358–1366, 2012.

Yuan, Lei, Liu, Jun, and Ye, Jieping. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pp. 352–360, 2011.

Zhao, S.D., Cai, T., and Li, H. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.