Learning Graphs with a Few Hubs

Rashish Tandon, Pradeep Ravikumar {RASHISH, PRADEEPR}@CS.UTEXAS.EDU
Department of Computer Science
The University of Texas at Austin, USA

Abstract

We consider the problem of recovering the graph structure of a "hub-networked" Ising model given i.i.d. samples, under high-dimensional settings, where number of nodes p could be potentially larger than the number of samples n. By a "hub-networked" graph, we mean a graph with a few "hub nodes" with very large degrees. State of the art estimators for Ising models have a sample complexity that scales polynomially with the maximum node-degree, and are thus ill-suited to recovering such graphs with a few hub nodes. Some recent proposals for specifically recovering hub graphical models do not come with theoretical guarantees, and even empirically provide limited improvements over vanilla Ising model estimators. Here, we show that under such low sample settings, instead of estimating "difficult" components such as hub-neighborhoods, we can use quantitative indicators of our inability to do so, and thereby identify hub-nodes. This simple procedure allows us to recover hub-networked graphs with very strong statistical guarantees even under very low sample settings.

1. Introduction

Graphical Models are a popular class of multivariate probability distributions that are widely used in applications across science and engineering. The key idea here is to represent probability distributions compactly as a product of functions over the cliques of an underlying graph. The task of graphical model selection is to learn the underlying undirected graph given samples drawn from the distribution it represents. This task becomes particularly difficult in high-dimensional data settings, where the number of variables p could be larger than the number of samples n.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Due in part to its importance, many practical algorithms with strong statistical guarantees have been proposed for this graphical model selection problem. In this paper, we focus on binary Ising models, where the variables are binary. For such Ising graphical models, Ravikumar et al. (2010) show that "local" node-wise ℓ_1-regularized logistic regressions can recover the underlying graph exactly with high probability, when given $n = O(d^3 \log(p))$ i.i.d. samples, where p is the number of nodes, and d is the maximum node-degree of the graph. Another class of methods are based on local search and thresholding (Abbeel et al., 2006; Csiszár & Talata, 2006; Bresler et al., 2008; Anandkumar et al., 2011), but in the absence of other stringent assumptions, their computational complexity scales exponentially with the local node-degrees d. Among more "global" approaches, Ji & Seymour (1996); Csiszár & Talata (2006); Peng et al. (2009) and others have proposed penalized pseudo-likelihood (Besag, 1975) based approaches; while Yang & Ravikumar (2011) have proposed penalized estimators based on variational approximations to the graphical model log-likelihood; however the sample complexity of these methods also scale polynomially with the maximum node-degree of the graph.

In this paper, we consider the setting where the graphs have a few *hub nodes*, which are highly connected nodes whose degree could scale as large as linearly in the number of nodes. An importance instance of this are *power-law graphs*, which occur ubiquitously in many real-world settings, and in which hub-nodes with large degrees are few in number but not non-existent, and their maximum node degree could be very large. Since the sample complexity of the state of the art methods listed above scale polynomially with the maximum node-degree, they would thus not be very suitable in recovering such power-law graphs with hub nodes. Motivated by this, there have been a few statistical estimators proposed that explicitly target power-law graphical model estimation. Liu & Ihler (2011) propose a novel *non-convex* regularization motivated by the power-law degree distribution, a convex variant of which was also considered in (Defazio & Caetano, 2012). While these methods did not provide theoretical guarantees, even their experimental results demonstrated limited improve-

ments in sample complexity over ℓ_1 regularization based methods. Peng et al. (2009) propose a pseudo-likelihood based procedure for learning discrete graphical models that minimizes the sum of *weighted* node-wise conditional log-likelihoods, where the node-wise weights could potentially be tuned to encourage power-law structure, but this was suggested as a heuristic. For the specific case of Gaussian graphical models, Hero & Rajaratnam (2012) provide an approach based on thresholding sample partial correlation matrices, and provide asymptotic expressions for false discovery rates under stringent weak dependence assumptions.

Consider the following leading question: what if we do not have enough samples to solve for the node-conditional distribution of a hub-node in an Ising model i.e. what if we we have less than $d_h^3 \log p$ samples, where d_h is the degree of the hub node? The estimators above that focus on the estimation of a hub-networked graphical model all focus in part on the estimation of such "difficult" sub-problems; so that they have a large sample complexity for estimating such hub-networked graphical models (Tandon & Ravikumar, 2013). In this paper, we propose to turn the problem on its head, and use our *inability* to estimate such difficult sub-problems given limited samples, to then turn around and be able to estimate the hub-network. To provide intuition for our strategy, consider a star-shaped graph, with one hub node, and the rest being spoke nodes connected only to the hub. The maximum degree of the hub node is thus $p-1$, so that estimating the node-conditional distribution of the hub-node would require samples scaling as $p^3 \log p$. What if only have samples scaling as $\log p$? But suppose we are also able to realize that we are *unable* to estimate the node-conditional distribution of the hub-node; and only those of the spoke nodes. We can then ignore the neighborhood estimation of the hub-node, and use the reliable neighborhood estimates of just the spoke nodes: this suffices to estimate the star-graph.

In this paper, we formalize this strategy: we provide a quantitative criterion for checking whether or not the given number of samples suffice for regularized node-conditional distribution estimation as in (Ravikumar et al., 2010) at a given node. We then use this to detect "hub nodes," and use only the neighborhood estimates from the remaining nodes to construct the graph estimate. We note that our notion of "hub nodes" is specifically related to the difficulty of node-neighborhood estimation, which only roughly corresponds to the node-degree (while the required sample size scales as $O(d^3 \log p)$, the constants matter in finite sample settings).

Our criterion is based on the following key observations on ℓ_1 regularized node-neighborhood estimation for any node $u \in V$ conditioned on the rest of the nodes. Consider the variance of the Bernoulli event of the incidence of any node $v \in V \backslash u$ in the node-neighborhood estimate, as a function the regularization penalty. When the penalty is very small, the node-neigborhood estimate will include all nodes, and the variance will be zero; when the penalty is "just right," the node-neighborhood estimate will be correct and will include v iff it is a neighbor with very high probability, so that the variance will again be (close to) zero, and when the penalty is very large, the node-neighborhood estimate will be null, and the variance will again be zero. Contrast this behavior with the setting where there are very few samples to allow for neighborhood recovery at any value of the regularization penalty: then the variance starts off at zero, rises, and then slowly goes to zero as the node-neighborhood becomes null. The difference in the observable behaviors between these two settings thus allows us to differentiate "hub" nodes from non-hubs. As we show, we are able to provide concrete statistical guarantees for our procedure, demonstrating improved sample complexity over the vanilla ℓ_1 regularized node-regression procedure.

We note that the approach of Liu & Ihler (2012) is similar in spirit to ours, utilizing a weighted combination of the node wise estimates to obtain the overall estimate, where the weights are the inverse of an alternate notion of variance. However, their approach deals with parameter estimation in the asymptotic sense, and is not applicable to structure estimation in the high dimensional setting.

Overall, our paper makes a key advance in the estimation of hub-networked graphical models: we provide a tractable procedure with strong statistical guarantees even under very low-sample settings where we cannot even estimate the node-conditional distributions of the hub nodes. Our methods involve binary reliability indicators for node-conditional distribution estimation, which could have broader applications in many scientific and engineering applications, even outside the context of graphical model estimation.

2. Notation and Preliminaries

Let $X = (X_1, \ldots, X_p)$ be a random vector, with each variable $X_i \, (i \in [p])$ taking values from a discrete set $\mathcal{X}$. Let $G = (V, E)$ be an undirected graph over p nodes, corresponding to the p variables $\{X_1, \ldots, X_p\}$. A pairwise Markov random field over $X = (X_1, \ldots, X_p)$ is a probability distribution specified by non-negative pairwise functions $\phi_{rt} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for each edge $(r, t) \in E$:

$$\mathbb{P}(x) \propto \prod_{rt \in E} \phi_{rt}(x_r, x_t) \tag{1}$$

Note that we use rt as a shorthand for the edge (r, t). In this paper, we focus on the Ising model setting *i.e.* where we have binary variables with $\mathcal{X} = \{-1, 1\}$, and where $\phi_{rt} = \exp(\theta_{rt} x_r x_t)$ for a given set of parameters $\theta =$

$\{\theta_{rt} \,|\, rt \in E\}$. In this case, (1) can be rewritten as :

$$\mathbb{P}_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{\sum_{rt\in E} \theta_{rt} x_r x_t\right\}, \tag{2}$$

where $Z(\theta) = \sum_{x\in\{-1,1\}^p} \exp\left\{\sum_{rt\in E} \theta_{rt} x_r x_t\right\}$.

Let $D := \{x^{(1)} \ldots, x^{(n)}\}$ be n samples drawn i.i.d from the Ising model distribution P_{θ^*} with parameters $\theta^* \in \mathbb{R}^{\binom{p}{2}}$ and Markov graph $G^* = (V, E^*)$, $|V| = p$. Note that each sample $x^{(i)}$ is a p-dimensional binary vector $x^{(i)} \in \{-1, 1\}^p$. The edge set E^* is related to the parameters θ^* as $E^* = \{(r, t) \in V \times V \,|\, \theta^*_{rt} \neq 0\}$.

The task of graphical model selection is to infer this edge set E^* using the n samples. Any estimator $\widehat{E}_n$ for this task is said to be sparsistent if it satisfies $\mathbb{P}\left[\widehat{E}_n = E^*\right] \to 1$ as $n \to \infty$.

2.1. ℓ_1-regularized estimator

We now briefly review the state-of-the-art estimator of (Ravikumar et al., 2010) (called the ℓ_1-estimator henceforth). The key idea there is to estimate the true graph E^* by estimating the neighbourhood of each node $r \in V$ in turn. Suppose $\mathcal{N}^*(r)$ denotes the true neighbours of the vertex r, so that $\mathcal{N}^*(r) = \{t \,|\, (r, t) \in E^*\}$. The ℓ_1-estimator uses sparsistent neighborhood estimators $\widehat{\mathcal{N}}_n(r) \subset V \; \forall r \in V$ s.t. $\mathbb{P}\left[\widehat{\mathcal{N}}_n(r) = \mathcal{N}^*(r)\right] \to 1$ as $n \to \infty$, to then obtain a sparsistent estimate of the entire graph.

Note that for any $r \in V$, the set of parameters θ^* is related to the true neighbourhood as $\mathcal{N}^*(r) = \{t \,|\, \theta^*_{rt} \neq 0, t \in V\}$. The ℓ_1-estimator exploits this to pose neighbourhood selection as an ℓ_1-regularized logistic regression problem, minimizing the negative conditional log-likelihood for each node with an additional ℓ_1-penalty. Note that for a set of parameters θ and a node $r \in V$, the conditional distribution of X_r conditioned on $X_{V\setminus r}$ is given as

$$\mathbb{P}_\theta\left(x_r \,|\, x_{V\setminus r}\right) = \frac{\exp(2x_r \sum_{t\in V\setminus r} \theta_{rt} x_t)}{1 + \exp(2x_r \sum_{t\in V\setminus r} \theta_{rt} x_t)}. \tag{3}$$

Defining $\theta_{\setminus r} = \{\theta_{rt} \,|\, t \in V, t \neq r\}$ and $x_{\setminus r} = \{x_t \,|\, t \in V, t \neq r\}$, the negative conditional log-likelihood of the samples D would be given by

$$\mathcal{L}(\theta_{\setminus r}; D) = \frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(1 + \exp\left(2x_r^{(i)}\theta_{\setminus r}^T x_{\setminus r}^{(i)}\right)\right) - 2x_r^{(i)}\theta_{\setminus r}^T x_{\setminus r}^{(i)}\right\}. \tag{4}$$

The ℓ_1-estimator solves the following optimization problem for each $r \in V$:

$$\underset{\theta_{\setminus r}\in\mathbb{R}^{p-1}}{\arg\min}\left\{\mathcal{L}(\theta_{\setminus r}; D) + \lambda\|\theta_{\setminus r}\|_1\right\}. \tag{5}$$

Let $\widehat{\theta}_{\setminus r}(D)$ correspond to the solution of (5). Then the neighbourhood estimate is given as the non-zero locations or support of $\widehat{\theta}_{\setminus r}(D)$: $\widehat{\mathcal{N}}_\lambda(r; D) = \text{Support}\left(\widehat{\theta}_{\setminus r}(D)\right)$. Finally, the edge estimate is computed by taking the union of all neighbourhood estimates: $\widehat{E}_{n,\lambda} = \bigcup_{r\in V}\{(r, t) \,|\, t \in \widehat{\mathcal{N}}_\lambda(r; D)\}$.

The ℓ_1-estimator has been shown to have strong statistical guarantees under certain incoherence conditions. Below, we restate the incoherence conditions of (Ravikumar et al., 2010), for the sake of completeness. These are stated in terms of the Hessian (in expectation) of the likelihood function for the true parameter vector $\theta^*_{\setminus r}$, which is given as $Q^*_r = \mathbb{E}\left[\nabla^2 \log P_{\theta^*}\left(x_r \,|\, x_{V\setminus r}\right)\right]$. For brevity, we shall briefly write Q^*_r as Q^*, the true neighbourhood set $\mathcal{N}^*(r)$ as $\mathcal{N}$, and its complement, $V \setminus \mathcal{N}^*(r)$ as $\mathcal{N}^c$. Then, their incoherence conditions (with $r \in V$ being implicit in Q^* and $\mathcal{N}$) are :

(A1) $\exists$ a const. $C_{\min} > 0$ s.t. $\Lambda_{\min}\left(Q^*_{\mathcal{N}\mathcal{N}}\right) \geq C_{\min}$. Also, $\exists$ a const. $C_{\max}$ s.t. $\Lambda_{\max}\left(\mathbb{E}\left[X_{V\setminus r} X_{V\setminus r}^T\right]\right) \leq C_{\max}$

(A2) $\exists$ a constant $\alpha \in (0, 1]$ s.t. $\left\|Q^*_{\mathcal{N}^c\mathcal{N}}\left(Q^*_{\mathcal{N}\mathcal{N}}\right)^{-1}\right\|_\infty \leq 1 - \alpha$

Note that $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ correspond to the minimum and maximum eigenvalues of a matrix respectively, and $\|\cdot\|_\infty$ corresponds to the standard ℓ_∞-matrix norm.

Now, we restate the main theorem below from (Ravikumar et al., 2010) using our notation, and refer the reader to their paper for details.

Theorem 1 (Guarantee for the ℓ_1-estimator; see (Ravikumar et al., 2010))**.** *Suppose an Ising graphical model with true parameter set θ^* satisfies conditions (A1) and (A2) for all nodes $r \in V$. Consider any $r \in V$, and let $d_r = \|\theta^*_{\setminus r}\|_0$ denote its degree. Then, there exist constants c_1, c_2, c_3, c_4 such that if we have $\lambda \geq c_1\sqrt{\frac{\log p}{n}}$ and $n > c_2\, d_r^3 \log p$ and $\mathcal{N}^*_{sub}(r) = \left\{t \in \mathcal{N}^*(r) \,\middle|\, |\theta^*_{rt}| \geq c_3\sqrt{d_r}\lambda\right\}$, then*

$$\mathbb{P}\left(\widehat{\mathcal{N}}_\lambda(r; D) = \mathcal{N}^*_{sub}(r)\right) \geq 1 - 2\exp\left(-c_4\lambda^2 n\right). \tag{6}$$

Based on Theorem 1, and a simple application of the union bound, we can see that the sample complexity for recovering the entire graph scales as $n = \Omega(d^3_{\max} \log p)$ samples, where $d_{\max}$ is the maximum degree of the graph

$G^* = (V, E^*)$. However, as detailed earlier, $d_{\max}$ may be huge for hub-graphs, so that the sample complexity of the ℓ_1-estimator will be large for such graphs.

3. Our Algorithm

As noted in the introduction, our approach is based on using a quantitative criterion for checking whether or not the given number of samples suffice for regularized node-conditional distribution estimation as in the ℓ_1-estimator at a given node. Given such a criterion, we can then take the union of only those neighborhood estimates which the method is guaranteed to estimate accurately, and not consider the "junk" estimates. Towards building such a observable "sufficiency" criterion, we first setup some notation.

3.1. Sufficiency Measure

For every $r \in V$ and $t \in V \setminus r$, we define $p_{r,n,\lambda}(t) = \mathbb{P}\left(t \in \widehat{\mathcal{N}}_\lambda(r; D)\right)$, as the probability of variable t being included in the neighbourhood estimate of variable r, estimated by the ℓ_1-estimator with regularization λ, given n samples drawn i.i.d. from the underlying Ising model. Note that the probability is taken over n samples. Based on Theorem 1, we have the following simple corollary.

Corollary 1. *For any $r \in V$, suppose θ^* and (n, λ) satisfy all conditions of Theorem 1 with constants c_1, c_2, c_3, c_4; then*

$$\begin{aligned} &p_{r,n,\lambda}(t) \geq 1 - 2\exp\left(-c_4\lambda^2 n\right) \quad \text{if } t \in \mathcal{N}^*_{sub}(r) \text{ and,} \\ &p_{r,n,\lambda}(t) \leq 2\exp\left(-c_4\lambda^2 n\right) \quad \text{if } t \notin \mathcal{N}^*_{sub}(r), \end{aligned} \tag{7}$$

where $\mathcal{N}^*_{sub}(r) = \left\{t \in \mathcal{N}^*(r) \,\middle|\, |\theta^*_{rt}| \geq c_3\sqrt{d}\lambda\right\}$.

Thus, when the number of samples n is sufficient for neighborhood recovery, depending on whether node t is in the true neighborhood of r, $p_{r,n,\lambda}(t)$ goes extremely close to zero or one; equivalently $p_{r,n,\lambda}(t)\left(1 - p_{r,n,\lambda}(t)\right)$ goes extremely close to zero. Building on this observation, let us define the "sufficiency" measure

$$\mathcal{M}_{r,n,\lambda} = \max_{t \in V\setminus r} p_{r,n,\lambda}(t)\left(1 - p_{r,n,\lambda}(t)\right). \tag{8}$$

It can thus be seen that this sufficiency measure goes to zero when the number of samples n is sufficient for recovering the neighborhood of node r.

In the sequel, we will analyze a natural U-statistic to estimate this sufficiency measure from data. We first require some more notation. For any b $\left(1 < b < \frac{n}{2}\right)$, we define $S_b(D)$ as the set of all possible subsamples of size b, drawn from D without replacement, so that

$$S_b(D) = \{(x^{(i_1)}, \ldots, x^{(i_b)}) \,|\, 1 \leq i_1 < \ldots < i_b \leq n\}. \tag{9}$$

Given any subsample $D_b \in S_b(D)$ of size b, let $F^t_{\lambda,r}(D_b)$ be a function such that

$$F^t_{\lambda,r}(D_b) = \begin{cases} 1 & \text{if } t \in \widehat{\mathcal{N}}_{b,\lambda}(r; D_b) \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Now, we consider the U-statistic (of order b),

$$\widetilde{p}_{r,b,\lambda}(t; D) = \frac{1}{\binom{n}{b}} \sum_{D_b \in S_b(D)} F^t_{\lambda,r}(D_b). \tag{11}$$

Note that $\mathbb{E}\left[\widetilde{p}_{r,b,\lambda}(t; D)\right] = p_{r,b,\lambda}(t)$. We are now ready to provide the U-statistic estimate of the sufficiency measure in (8):

$$\widetilde{\mathcal{M}}_{r,b,\lambda}(D) = \max_{t \in V\setminus r} \widetilde{p}_{r,b,\lambda}(t; D)\left(1 - \widetilde{p}_{r,b,\lambda}(t; D)\right). \tag{12}$$

Computing $\widetilde{\mathcal{M}}_{r,b,\lambda}(D)$ would require computing $\widetilde{p}_{r,b,\lambda}(t; D)$ for every $t \in V \setminus r$, which in turn would require considering all possible $\binom{n}{b}$ sub-samples of D. However, as we show below (see also analyses in (Liu et al., 2010; Politis et al., 1999) on sub-sampling), it suffices to choose a number $N \geq n/b$ of subsamples drawn at random. Thus, our actual estimate for $p_{r,b,\lambda}(t)$ is

$$\widehat{p}_{r,b,\lambda}(t; D) = \frac{1}{N} \sum_{i=1}^{N} F^t_{\lambda,r}(D_i), \tag{13}$$

where $D_1, \ldots, D_N$ are subsamples chosen independently and uniformly at random from $S_b(D)$, and the estimate for the sufficiency measure is

$$\widehat{\mathcal{M}}_{r,b,\lambda}(D) = \max_{t \in V\setminus r} \widehat{p}_{r,b,\lambda}(t; D)\left(1 - \widehat{p}_{r,b,\lambda}(t; D)\right). \tag{14}$$

We describe the procedure to calculate $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$ in Algorithm 1.

Once $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$ has been computed, we have the following lemma which shows that it is ϵ-close to $\mathcal{M}_{r,b,\lambda}$ with high probability, provided we have sufficiently many samples .

Proposition 1 (Concentration of $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$ to $\mathcal{M}_{r,b,\lambda}$)**.** *For any $\delta \in (0, 1]$ and $\epsilon > 0$, if we have $n > \frac{18b}{\epsilon^2}\left[\log p + \log\left(4/\delta\right)\right]$ and $N \geq \lceil \frac{n}{b} \rceil$, then,*

$$\mathbb{P}\left(|\widehat{\mathcal{M}}_{r,b,\lambda}(D) - \mathcal{M}_{r,b,\lambda}| \leq \epsilon\right) \geq 1 - \delta. \tag{15}$$

3.2. Behavior of the Sufficiency Measure

The key question of interest is whether we can use the sufficiency measure $\mathcal{M}_{r,b,\lambda}$ (via its sample estimate

Algorithm 1 Estimating $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$

Input : Data $D := \{x^{(1)}, \ldots, x^{(n)}\}$, Regularization parameter λ, Sub-sample size b , No. of sub-samples N

Output: An estimate of $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$

$\forall\, t \in V \setminus r, \quad \widehat{p}_{r,b,\lambda}(t; D) \leftarrow 0$

for $i = 1$ **to** N **do**

  Pick a sub-sample D_i chosen uniformly randomly from $S_b(D)$

  Compute $\widehat{\mathcal{N}}_{b,\lambda}(r; D_i)$ by solving (5) (ℓ_1-estimate)

  for $t \in \widehat{\mathcal{N}}_{b,\lambda}(r; D_i)$ **do**

    $\widehat{p}_{r,b,\lambda}(t; D) \leftarrow \widehat{p}_{r,b,\lambda}(t; D) + 1$

$\forall\, t \in V \setminus r, \quad \widehat{p}_{r,b,\lambda}(t; D) \leftarrow \widehat{p}_{r,b,\lambda}(t; D)/N$

$$\widehat{\mathcal{M}}_{r,b,\lambda}(D) \leftarrow \max_{t \in V \setminus r} \widehat{p}_{r,b,\lambda}(t; D)\,(1 - \widehat{p}_{r,b,\lambda}(t; D))$$

$\widehat{\mathcal{M}}_{r,b,\lambda}(D)$) to detect "hub-nodes" that we define specifically as those nodes for which we do not have enough samples for the ℓ_1-estimator to be sparsistent. Correspondingly, let us define "non-hub" nodes in this context as those nodes for which we do have enough samples for the ℓ_1-estimator to be sparsistent. We formalize these notions below.

Definition 1 (Non-Hub Node vs. Hub Node)**.** *Assume that the true parameter set θ^* satisfies the incoherence conditions, $(A1)$ and $(A2)$, for all nodes $r \in V$. Consider any node $r \in V$. It is termed a "non-hub node" w.r.t. n samples if $\exists$ a regularization parameter λ s.t. (n, λ) satisfy all conditions of Theorem 1 with constants c_1, c_2, c_3, c_4. Otherwise, the node is termed a "hub" node.*

Since the sample complexity of neighborhood estimation via the ℓ_1-estimator scales cubically with the node-degree (from Theorem 1), hub nodes as we define here correspond loosely to high-degree nodes, but in the sequel, the exact specification of "hub" and "non-hub" nodes are as detailed by the definition above.

Before we describe the behaviour of $\mathcal{M}_{r,b,\lambda}$ for "hub" nodes and "non-hub" nodes, we impose the following technical assumptions on the behaviour of $p_{r,b,\lambda}(t)$ needed for our algorithm to work.

Assumption 1. $\forall\, r \in V$, $p_{r,b,\lambda}(t)$ *satisfies the following: For fixed b and some constant $c(> 0)$, let*

$$\lambda_{\min}(t) = \min\left\{\lambda \geq 0 \,|\, p_{r,b,\lambda}(t) \leq 1 - 2\exp\left(-c \log p\right)\right\},$$

and,

$$\lambda_{\max}(t) = \max\left\{\lambda \geq 0 \,|\, p_{r,b,\lambda}(t) \geq 2\exp\left(-c \log p\right)\right\}. \tag{16}$$

Then, $\lambda_{\min}(t)$ and $\lambda_{\max}(t)$ are attained at finite values s.t.

(a) For any $t \in V \setminus r$ and $\lambda \in (\lambda_{\min}(t), \lambda_{\max}(t))$, we have

$$p_{r,b,\lambda}(t) \in [2\exp(-c \log p), 1 - 2\exp(-c \log p)]\,. \tag{17}$$

(b) For all $t \notin \mathcal{N}^*(r)$,

$$\lambda_{\min}(t) \leq \lambda_{\min} < \lambda_{\max} \leq \lambda_{\max}(t), \tag{18}$$

for some finite $\lambda_{\min}, \lambda_{\max} \geq 0$ independent of t.

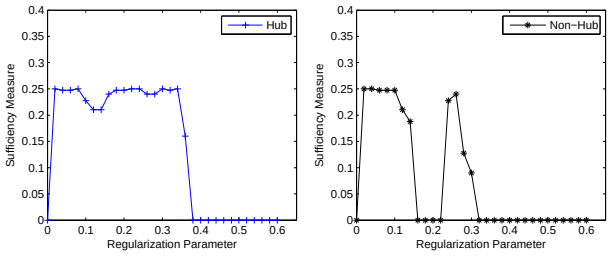
(c) For any $t \in V \backslash r$, $\exists\, t' \notin \mathcal{N}^*(r) : \lambda_{\min}(t') < \lambda_{\max}(t)$.

Additionally, $p_{r,b,\lambda}(t)$ is a continuous function of λ.

To build intuition for the assumptions, as well as our analysis in the sequel, it will be instructive to consider the behavior of the inclusion probability $p_{r,b,\lambda}(t)$ as we increase λ from zero to infinity. When λ is zero, the ℓ_1-estimator reduces to the unregularized conditional MLE: any variable $t \in V \backslash r$ will always occur in the neighborhood estimate of node r, and $p_{r,b,\lambda}(t)$ will be equal to one. As λ increases, the inclusion probability in turn reduces, and at a very large value of λ, the inclusion probability $p_{r,b,\lambda}(t)$ will become equal to zero: this follows from the property of the ℓ_1-estimator, where there exists a large regularization weight when the parameter estimate becomes equal to zero.

In the assumptions above, it can be seen that if $\lambda_{\min}(t)$ and $\lambda_{\max}(t)$ exist, then by definition, we must have $\lambda_{\min}(t) \leq \lambda_{\max}(t)$. Part (a) of the assumption is a smoothness constraint that ensures that if the probability of inclusion or exclusion of a variable into a neighbourhood gets close to 1, then it stays close to 1, and does not vary wildly. Part (b) ensures that ranges of $[\lambda_{\min}(t), \lambda_{\max}(t)]$ intersect at least for all irrelevant variables $t \notin \mathcal{N}^*(r)$. This is a very mild assumption that ensures that the inclusion probability of an irrelevant variable does not stay exactly at one as we increase λ, and reduces at least very slightly (below the threshold of $1 - 2\exp(-c \log p)$) before other irrelevant variables have their inclusion probability drop from one all the way to zero. Part (c) is a closely related mild assumption that ensures that the probability of inclusion of atleast one irrelevant variable would have dropped by a small value from 1 before any other variable has its inclusion probability drop from one all the way to zero. We note that these mild technical assumptions on the inclusion probabilities always hold in our empirical observations.

Armed with these assumptions, we now analyze the behavior of our sufficiency measure $\mathcal{M}_{r,b,\lambda}$. Our next proposition shows that there exists atleast one "bump" in the graph of the sufficiency measure against the regularization penalty λ.



(a) $\mathcal{M}_{r,b,\lambda}$ for a hub node ($degree = 19$) (b) $\mathcal{M}_{r,b,\lambda}$ for a non-hub node ($degree = 1$)

Figure 1. Behaviour of $\mathcal{M}_{r,b,\lambda}$ for non-hub nodes and hub-nodes in a star graph on $p = 100$ nodes

Proposition 2 ("First Bump"). *Suppose Assumption 1 holds with constant $c > 0$. Let*

$$\gamma = 2\exp(-c\log p)\left(1 - 2\exp(-c\log p)\right). \tag{19}$$

For any node $r \in V$, let $\lambda_l = \inf\{\lambda \geq 0 : \mathcal{M}_{r,b,\lambda} \geq \gamma\}$ be the smallest regularization penalty where the sufficiency measure is greater than a small threshold above zero, and $\lambda_u = \inf\{\lambda > \lambda_l : \mathcal{M}_{r,b,\lambda} < \gamma\}$ be the next value of the regularization penalty where the sufficiency measure falls below that threshold. Then, (a) the infima above are attained at finite values, and (b) for any $k \in (\gamma, 1/4]$, $\exists \lambda \in (\lambda_l, \lambda_u)$ s.t. $\mathcal{M}_{r,b,\lambda} \geq k$.

Our next two propositions track the behavior of the ℓ_1-estimate $\widehat{\mathcal{N}}_{b,\lambda}(r; D)$ after the first bump outlined above. The very next proposition provides the behavior for "non-hub" nodes.

Proposition 3 (Behavior at λ_u for "non-hub nodes"). *Let $r \in V$ be a "non-hub node" w.r.t. b samples, and constants for all conditions of Theorem 1 being c_1, c_2, c_3, c_4. Let Assumption 1 hold for a constant $c > 1$ with $c < c_1 c_4$, and let λ_u be as defined in Proposition 2. Then, $\widehat{\mathcal{N}}_{b,\lambda_u}(r; D)$ recovers the neighborhood with high probability:*

$$\mathbb{P}\left(\mathcal{N}^*_{sub}(r) \subseteq \widehat{\mathcal{N}}_{b,\lambda_u}(r; D) \subseteq \mathcal{N}^*(r)\right) > 1 - 2\exp\left(-(c-1)\log p\right),$$

where $\mathcal{N}^*_{sub}(r) = \left\{t \in \mathcal{N}^*(r) \,\middle|\, |\theta^*_{rt}| \geq c_3\sqrt{d}\lambda\right\}$.

The proposition thus tells us that for "non-hub nodes", after the first bump when the value of $\mathcal{M}_{r,b,\lambda}$ becomes very close to zero, the ℓ_1-estimator recovers $\mathcal{N}^*_{\text{sub}}(r)$ w.h.p. (as also indicated by Theorem 1). Note that when increasing λ further, there would be further bump(s): the value of $\mathcal{M}_{r,b,\lambda}$ would rise, but would again drop back to zero: when λ is very large, the neighborhood estimate is null, so that the probability for any node to be in the neighborhood will be exactly zero; so that the sufficiency measure will be equal to zero. Figure 1(b) demonstrates this behavior in a simulated dataset.

On the other hand, for "hub nodes", the behavior of $\widehat{\mathcal{N}}_{b,\lambda}(r; D)$ at $\lambda = \lambda_u$, defined in Proposition 2, is given by the following proposition.

Proposition 4 (Behavior at λ_u for "hub nodes"). *Let $r \in V$ be a "hub node" w.r.t. b samples. Also, let Assumption 1 hold with constant $c > 1$. Then $\widehat{\mathcal{N}}_{b,\lambda_u}(r; D)$ excludes irrelevant variables with high-probability:*

$$\mathbb{P}\left(\widehat{\mathcal{N}}_{b,\lambda_u}(r; D) \subseteq \mathcal{N}^*(r)\right) > 1 - 2\exp\left(-(c-1)\log p\right).$$

The proposition thus tells us that for "hub nodes", after the first bump when the value of $\mathcal{M}_{r,b,\lambda}$ becomes very close to zero, irrelevant variables are excluded, though however there is no guarantee on relevant variables being included. Empirically in fact, the end of the first bump typically occurs at a very large value of λ when *all* variables are excluded; in particular, the graph of $\mathcal{M}_{r,b,\lambda}$ against λ typically has a single bump. Figure 1(a) demonstrates this behavior in a simulated dataset.

Propositions 3 and 4 thus motivate using the behaviors of the sufficiency measure as outlined above to distinguish hub nodes and non-hub nodes; and then compute the graph estimate using the neighborhood estimates from the non-hubs alone. This natural procedure is described in Algorithm 2.

Algorithm 2 Algorithm to compute neighborhood estimate $\widehat{\mathcal{N}}(r)$, for each node $r \in V$, and the overall edge estimate $\widehat{E}$

Input : Data $D := \{x^{(1)}, \ldots, x^{(n)}\}$, Regularization parameters $\Lambda := \{\lambda_1, \ldots, \lambda_s\}$, Sub-sample size b, No. of sub-samples N, Thresholds on sufficiency measure t_l and t_u, Node $r \in V$

Output: An estimate $\widehat{\mathcal{N}}(r)$ of the neighborhood for each $r \in V$, and the overall edge estimate $\widehat{E}$

foreach $r \in V$ **do**

- $\forall\, \lambda \in \Lambda$, Compute $\widehat{\mathcal{M}}_{r,b,\lambda}(D)$ using Algorithm 1
- $\lambda' \leftarrow$ Smallest $\lambda \in \Lambda$ s.t. $\widehat{\mathcal{M}}_{r,b,\lambda}(D) > t_u$
- $\Lambda \leftarrow \{\lambda \in \Lambda : \lambda > \lambda'\}$
- $\lambda_0 \leftarrow$ Smallest $\lambda \in \Lambda$ s.t. $\widehat{\mathcal{M}}_{r,b,\lambda}(D) < t_l$
- $\widehat{\mathcal{N}}(r) \leftarrow \left\{t \,|\, \widehat{p}_{r,b,\lambda_0}(t; D) \geq \frac{1+\sqrt{1-4t_l}}{2}\right\}$

$\widehat{E} \leftarrow \bigcup_{r \in V} \{(r, t) \,|\, t \in \widehat{\mathcal{N}}(r)\}$

The following theorem is a natural corollary of Theorem 1, and Propositions 3 and 4. Note that in the below, we assume that the true parameter set θ^* satisfies the incoherence conditions, $(A1)$ and $(A2)$, for all nodes $r \in V$; and that

Assumption 1 holds $\forall\, r \in V$, with an appropriate constant $c > 2$, satisfying conditions of Proposition 3 for "non-hub nodes".

Theorem 2 (Guarantee for the estimator of Algorithm 2)**.** *Suppose we run Algorithm 2 setting* $t_l = 2\exp(-c\log p)(1 - 2\exp(-c\log p)) + \epsilon$, $t_u = 1/4 - \epsilon$, *the sub-sample size* $b = f(n)$ *(with* $\sqrt{n} \leq f(n) < n/2$*), and number of sub-samples* $N \geq \lceil n/f(n) \rceil$, *such that*

$$n > 18 f(n) \left[\log p + \log\left(4/\delta\right)\right] / \epsilon^2. \tag{20}$$

For any degree-value $d \in \{1, \ldots, p\}$ *and constant* $c'' > 0$, *denote*

$$E_d = \left\{ (s,t) \in E^* \;\middle|\; \min(d(s), d(t)) \leq d,\, |\theta^*_{st}| \geq c'' \sqrt{\frac{d \log p}{n}} \right\} \tag{21}$$

where $d(v)$ *corresponds to the degree of vertex* v *in* E^*. *Then, there exist constants* c, c', c'', c''', *such that if the sub-sample size scales as*

$$f(n) > c' d^3 \log p, \tag{22}$$

then the graph structure estimate $\widehat{E}$ *of Algorithm 2 satisfies:*

$$\mathbb{P}\left(E_d \subseteq \widehat{E} \subseteq E^*\right) \geq 1 - 2\exp\left(-c''' \log p\right) - \delta. \tag{23}$$

Now, let us define the *critical degree*, d_c, of a graph $G^* = (V, E^*)$, as the minimum degree such that neighborhoods of vertices with at most the said degree cover the whole graph, *i.e.*

$$\begin{aligned} d_c = \min \quad & d \\ \text{s.t. } \forall\, (s,t) \in E^*, \; & \text{either } d(s) \leq d \text{ or } d(t) \leq d. \end{aligned} \tag{24}$$

The following corollary then gives the sample complexity for exact recovery of the graph, assuming that the edges have sufficient weight.

Corollary 2. *Let the conditions of Theorem 2 be satisfied, with* $b = f(n)$ *as the sub-sample size. Let* d_c *be the critical degree of the graph* G^*. *Then there exist constants* c', c'', c''' *s.t. if the sub-sample size scales as*

$$f(n) > c' d_c^3 \log p, \tag{25}$$

and $|\theta^*_{st}| \geq c'' \sqrt{\frac{d_c \log p}{n}}\, \forall\, (s,t,) \in E^*$, *then*

$$\mathbb{P}\left(\widehat{E} = E^*\right) \geq 1 - 2\exp(-c''' \log p), \tag{26}$$

where $\widehat{E}$ *is the graph structure estimate from Algorithm 2.*

Note that we may choose $f(n) = c' n^{1-\rho}$, for some value of $\rho \in (0, 0.5]$, as the sub-sample size. The choice of ρ would be governed by d_c for the graph under consideration. For example, if d_c is a constant (e.g. $d_c = 1$ in a star graph), then the optimal choice of ρ would be 0.5, yielding a overall sample complexity of $\Omega\left((\log p)^2\right)$.

4. Experiments

In this section, we present experimental results demonstrating that our algorithm does indeed succeed in recovering graphs with a few hubs.

4.1. Synthetic Data

We first performed structure learning experiments on simulated data using 3 types of graphs:

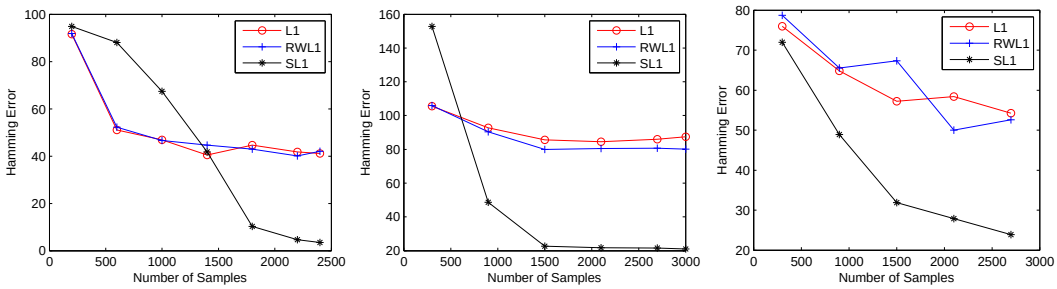
(a) a collection of stars with $p = 100$ nodes involving 5 hub nodes with degree $d = 19$, each connected to 19 other degree $d = 1$ nodes.

(b) a grid graph with 81 nodes (9×9), with 2 additional high degree hub-nodes of degree $d = 12$ (so that $p = 83$) attached to random points in the grid.

(c) a power-law graph on $p = 100$ nodes generated using the preferential attachment scheme (Barabasi & Albert, 1999).

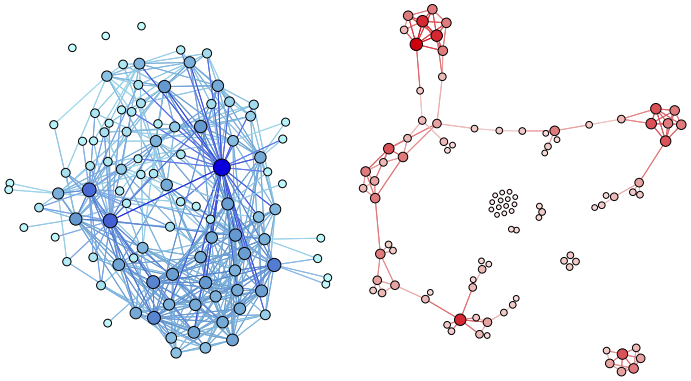
For each graph, we considered a pairwise Ising model with edge weight $\theta^*_{rt} = \frac{\omega}{\max(d_r, d_t)}$, for some $\omega > 0$, and where d_r and d_t were the degrees of r and t respectively. For each such Ising model, we generated n i.i.d. samples $D = \{x^{(1)}, \ldots, x^{(n)}\}$ using Gibbs sampling.

In all our experiments, for our algorithm (denoted as SL1 in our plots), the value of N, the number of times to subsample, was fixed to 60. We set lower and upper thresholds on the sufficiency measure as $t_l = 0.1$ and $t_u = 0.2$. The number of subsamples was set to $b = \min\left(20\sqrt{n}, \frac{n}{2}\right)$ and the set of regularization parameters was taken as $\Lambda = \{0.005, 0.01, 0.015, \ldots, 1\}$. We performed comparisons with the ℓ_1-estimator (Ravikumar et al., 2010) (denoted as L1 in our plots) and the reweighted ℓ_1-estimator for scale-free graphs (Liu & Ihler, 2011) (denoted as RWL1 in our plots). For both these methods, the best regularization parameter was chosen using the Bayesian information criterion (BIC) from the grid of regularization parameters Λ. Figure 2 shows plots of the Average Hamming Error (*i.e.* average number of mismatches from the true graph) with varying number of samples for our method and the baselines, computed over an average of 10 trials. Since our estimate uses subsamples to compute its sufficiency measure, when the number of samples is extremely low, the deviation of the sample sufficiency measure estimate from the population sufficiency measure becomes large enough so that the resulting mistakes made by our method in designating hubs and non-hubs increase its overall Hamming error. We note however that at such extremely low number of samples, it can be seen that the overall Hamming error of any estimator is quite high, so that none of the estimators provide useful graph estimates in any case. It can be seen



(a) Stars Graph (b) Hub+Grid Graph (c) Preferential Attachment Graph

Figure 2. Plots of Average Hamming Error vs Number of Samples



(a) L1 with BIC (b) Our Algorithm

Figure 3. Graphs obtained using US Senate voting records data from the 109th congress (Bannerjee et al., 2008)

that other than at such extremely few samples, we achieve *much lower* Hamming error than both L1 and RWL1, and which is particularly pronounced for scale-free graphs such as those provided by the preferential attachment model.

4.2. Real World Data

We ran our algorithm on a data set consisting of US senate voting records data from the 109th congress (2004 - 2006) (Bannerjee et al., 2008). It consists of 100 nodes ($p = 100$), corresponding to 100 senators. There are 542 samples, each representing a bill that was put to vote. For each *(senator, bill)* pair, the vote is recorded as either a 1 (representing a *yes*), a -1 (representing a *no*) or a 0 (representing a *missed vote*). For the purpose of the experiment, all 0 entries were replaced by -1, as also done in (Bannerjee et al., 2008).

Our algorithm was run with the parameters $N = 60, t_l = 0.1, t_u = 0.2, b = 450$ and $\Lambda = \{0.005, 0.01, 0.015, \ldots, 1\}$. Figure 3(b) shows the graph obtained using our method, whiles Figure 3(a) shows the graph obtained by running the ℓ_1-estimator (Ravikumar et al., 2010) with the regularization parameter being chosen using the Bayesian Information Criterion (BIC) from the set of regularization parameters Λ.

We see that the graph obtained using the ℓ_1-estimator with BIC is much denser than what we obtain. This also corroborates the observation of (Liu et al., 2010), that BIC leads to larger density in high dimensions. A few of the nodes in the graph using our algorithm are seen to have 0 degree, and are thus disconnected from the graph. This might be because these might be higher degree "hub" nodes, but for which the number of samples is not sufficient enough to provide a reliable estimate of the neighbourhoods vis-à-vis their degree. Overall, the sparse graph we obtained using our reliability indicator based method suggests the need for such reliability indicators to prevent the inclusion of spurious edge-associations.

Acknowledgements We acknowledge support from NSF via IIS-1149803 and DMS-1264033, and ARO via W911NF-12-1-0390.

References

Abbeel, P., Koller, D., and Ng, A. Y. Learning factor graphs in polynomial time and sample complexity. *Jour. Mach. Learning Res.*, 7:1743–1788, 2006.

Anandkumar, A., Tan, V. Y. F., and Willsky, A.S. High-Dimensional Structure Learning of Ising Models : Local Separation Criterion. *Preprint*, June 2011.

Bannerjee, O., , Ghaoui, L. El, and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Jour. Mach. Lear. Res.*, 9:485–516, March 2008.

Barabasi, A.L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Besag, J. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

Bresler, Guy, Mossel, Elchanan, and Sly, Allan. Reconstruction of markov random fields from samples: Some observations and algorithms. In *Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop, RANDOM 2008 on Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques*, APPROX '08 / RANDOM '08, pp. 343–356. Springer-Verlag, 2008.

Csiszár, I. and Talata, Z. Consistent estimation of the basic neighborhood structure of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

Defazio, A. and Caetano, T. A convex formulation for learning scale-free networks via submodular relaxation. In *Advances in Neural Information Processing Systems 24*, 2012.

Hero, A. and Rajaratnam, B. Hub discovery in partial correlation graphical models. *IEEE Trans. on Information Theory*, 58(9):6064–6078, 2012.

Ji, C. and Seymour, L. A consistent model selection procedure for markov random fields based on penalized pseudolikelihood. *Ann. Appl. Probab.*, 6(2):423–443, 1996.

Liu, Han, Roeder, Kathryn, and Wasserman, Larry A. Stability approach to regularization selection (stars) for high dimensional graphical models. In *NIPS*, pp. 1432–1440, 2010.

Liu, Q. and Ihler, A. T. Learning scale free networks by reweighted l1 regularization. *Journal of Machine Learning Research - Proceedings Track*, 15:40–48, 2011.

Liu, Qiang and Ihler, Alexander. Distributed parameter estimation via pseudo-likelihood. In *International Conference on Machine Learning (ICML)*, pp. 1487–1494. July 2012.

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

Politis, D., Romano, J.P., and Wolf, M. *Subsampling*. Springer series in statistics. Springer Verlag, 1999. ISBN 9780387988542.

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

Tandon, R. and Ravikumar, P. On the difficulty of learning power law graphical models. In *In IEEE International Symposium on Information Theory (ISIT)*, 2013.

Yang, E. and Ravikumar, P. On the use of variational inference for learning discrete graphical models. In *International Conference on Machine learning (ICML)*, 28, 2011.