Learning Scale Free Networks by Reweighted ℓ_1 regularization

Qiang Liu
Department of Computer Science
University of California, Irvine

Alexander Ihler
Department of Computer Science
University of California, Irvine

Abstract

Methods for ℓ_1-type regularization have been widely used in Gaussian graphical model selection tasks to encourage sparse structures. However, often we would like to include more structural information than mere sparsity. In this work, we focus on learning so-called "scale-free" models, a common feature that appears in many real-work networks. We replace the ℓ_1 regularization with a power law regularization and optimize the objective function by a sequence of iteratively reweighted ℓ_1 regularization problems, where the regularization coefficients of nodes with high degree are reduced, encouraging the appearance of hubs with high degree. Our method can be easily adapted to improve *any* existing ℓ_1-based methods, such as graphical lasso, neighborhood selection, and JSRM when the underlying networks are believed to be scale free or have dominating hubs. We demonstrate in simulation that our method significantly outperforms the a baseline ℓ_1 method at learning scale-free networks and hub networks, and also illustrate its behavior on gene expression data.

1 Introduction

A *scale-free network* is a network whose degree distribution follows a power law. Scale-free networks have been empirically observed in a wide variety of systems, including protein and gene networks, publication citation networks, and many social networks; for examples, see Barabási and Albert (2002) and references therein. It has been shown that scale free networks can be generated by *preferential attachment* mechanisms, in which new nodes prefer to connect to nodes with high degree (Barabási and Albert, 1999). Perhaps the most notable characteristic of a scale-free network is the relative frequency of "hubs" – vertices with a degree that greatly exceeds the average. Identifying hubs is often a primary step of understanding the network structure, as hubs are thought to serve specific, perhaps critical purposes in their network. For example, one important task in bioinformatics is to identify the hub proteins and hub genes.

On the other hand, Gaussian Markov random fields (GMRFs) have been widely used to infer the structure of networks from data. In particular, suppose that $x = [x_1, \cdots, x_p]'$ follows a p-dimensional Gaussian distribution with zero mean and covariance Σ. Let $\Omega = \Sigma^{-1}$ be the precision matrix. It is well known that the precision matrix reflects the dependency structure of the GMRFs, since the (i, j) element of Ω is non-zero if and only if x_i and x_j are connected in the Markov random field, i.e., are conditionally dependent given the values of all other elements of x. The task of inferring network structure can be thus recast as the statistical problem of estimating the precision matrix Ω. However, a major difficulty in this task derives from the fact that the number of observed data points n is usually small compared with the dimensionality p, in which case the empirical covariance matrix will have significant noise. The naïve method of simply taking the empirical covariance matrix usually results in a fully connected graph, and thus does not indicate any structural independence within the network. Worse, in the increasingly common case that $n < p$ the empirical covariance matrix becomes singular.

Various methods have been developed to estimate the structure of graphical models through the use of ℓ_1 regularizations (e.g., Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Peng et al., 2009; Banerjee et al., 2008). Among them, the neighborhood selection method of Meinshausen and Bühlmann (2006) is perhaps the simplest: by noting that the (i, j) element of Ω is, up to a positive scalar, the regression coefficient of variable j in the regression of variable i against the

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

rest, they estimate a sparse graphical model by fitting a collection of lasso regression models for each x_i using the other variables $x_{\neg i} = \{x_j | j \neq i\}$ in turn. They showed that this method provides an asymptotically consistent estimator of the set of non-zero elements of Ω. Building upon this approach, Peng et al. (2009) proposed a joint sparse regression method to address the asymmetry of Meinshausen and Bühlmann (2006). This method is related to the recent symmetric lasso method (Friedman et al., 2010), which is formulated as a maximum pseudolikelihood estimator (Besag, 1974). A more systematic approach is to maximize the ℓ_1 penalized log-likelihood (e.g., Yuan and Lin, 2007; Banerjee et al., 2008). Friedman et al. (2007) proposed an efficient blockwise coordinate descend method, called *graphical lasso*, which maximizes the ℓ_1 penalized log-likelihood by iteratively solving a series of modified lasso regressions.

However, it has been observed in previous work that ℓ_1-based methods work poorly for scale-free networks, in which some nodes (called hubs) have extremely high degrees. Schäfer and Strimmer (2005) showed an experiment in which Meinshausen and Bühlmann (2006) failed to identify the hub-type structures in gene networks. A possible reason for this is that the ℓ_1 regularization imposes sparsity uniformly and independently on each node, without any preference for identifying hub-like structures. Theoretically, the consistency result in Meinshausen and Bühlmann (2006) relies on an sparsity assumption that restricts the maximal size of the neighborhoods of variables (see Assumption 3 in Meinshausen and Bühlmann (2006)), which is inconsistent with the existence of hubs. For graphical lasso, Ravikumar et al. (2008) shows that to ensure an elementwise bound $||\hat{\Omega} - \Omega^{\text{true}}||_\infty = \mathcal{O}(\sqrt{\frac{\log p}{n}})$ holds with high probability, one need a sample size of $n = \mathbf{\Omega}(d^2 \log p)$, where d is the maximum degree of the graph. In other words, by this scaling analysis the sample size needs to be polynomial in the graph degree, while only logarithmic in its size (dimensionality). This suggests that the "curse of degree" is even worse than the "curse of dimensionality". Ravikumar et al. (2010) shows that a similar difficulty occurs in high dimensional Ising model selection when using ℓ_1-regularized logistic regression.

In fact, it has been widely realized that a simple ℓ_1 sparsity prior often does not take full advantage of the prior information that we may hold about the underlying model, especially when the true network has some specific structural features. There has been considerable recent work on improving the lasso by incorporating various prior information of structures in addition to simple sparsity, including the group lasso (Yuan and Lin, 2006), simultaneous lasso (Turlach et al., 2005) and fused lasso (Tibshirani et al., 2005) in variable selection problems. Analogously, there are many works incorporating different structured prior information in Gaussian graphical model selection in various ways, for example Duchi et al. (2008); Schmidt et al. (2009); Marlin et al. (2009); Friedman et al. (2010); Peng et al. (2010); Marlin and Murphy (2009); and Jacob et al. (2009). In this work, we develop an efficient method to incorporate a preference for scale-free behavior into our prior for network inference. We show that our method can be recast as a sequence of iteratively reweighted ℓ_1 regularization problems, where the regularization coefficients of the nodes with higher degree are decreased, encouraging the creation of hubs. We demonstrate the performance of our algorithm on both simulated and gene expression data.

2 Learning sparse networks

Suppose $x = [x_1, x_2, \cdots, x_p]'$ is drawn from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$, where Σ is the $p \times p$ covariance matrix. Let $\Omega = \Sigma^{-1} = \{\omega_{ij}\}$ be the precision matrix. $X = [x^1, x^2, \cdots, x^n]$ is a collection of n observed data. The task is to estimate the set of non-zero elements of Ω, which corresponds to the edges of the Gaussian graphical model. Let us denote $\rho_{ij} = \text{corr}(x_i, x_j | x_{\neg ij})$ to be the partial correlation, where $x_{\neg ij}$ are the other elements of x besides x_i and x_j. A well know fact is that the multivariate Gaussian model can be presented in a self-regression form, $x_i = \sum_{i \neq j} \beta_{ij} x_j + \delta_i$, where $\beta_{ij} = -\frac{\omega_{ij}}{\omega_{ii}} = \rho_{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}$, and δ_i are Gaussian noise that are independent of $x_{\neg i}$: $\delta_i \sim \mathcal{N}(0, 1/\omega_{ii})$. We will use B to represent a matrix with zero diagonal and β_{ij} as the off-diagonal entries, and Φ a matrix with zero diagonal and ρ_{ij} in the off-diagonals. Note that the off-diagonals of Ω, B and Φ only differ up to a nonzero constant, and hence share the same non-zero pattern; one can predict the model structure by estimating the non-zero pattern of any one of them.

A body of research has been developed to estimate the non-zero pattern by applying ℓ_1 regularization to the precision matrix or partial correlation matrix. Meinshausen and Bühlmann (2006) proposed a simple neghborhood selection method (henceforth referred to as "MB") by regressing each variable w.r.t. to all the other variables with a ℓ_1 regularization:

$$\min_{\beta} \frac{1}{2}||x_i - \sum_{j \neq i} \beta_{ij} x_j||^2 + \lambda \sum_{j \neq i} |\beta_{ij}|,$$

in which the non-zero elements of $\{\beta_{ij} | j \neq i\}$ decide the neighborhood of x_i. This leads to n independent standard lasso problems, which can be solved by efficient algorithms such as LARS or coordinate descent

(Friedman et al., 2007). It was shown that it gives asymptotical consistent estimator of the set of non-zero elements of precision matrix. Its disadvantage, however, is that since β_{ij} and β_{ji} are estimated independently, it cannot guarantee that the non-zero pattern of β_{ij} is symmetric and must use post-processing to create a symmetric estimate.

Peng et al. (2009) proposed a joint regression sparse method ("JSRM") to address this asymmetry issue, by minizing the following joint loss function

$$\frac{1}{2}\sum_{i}^{p}\mu_i||x_i - \sum_{j\neq i}\rho_{ij}\sqrt{\frac{\omega_{jj}}{\omega_{ii}}}x_j||^2 + \lambda\sum_{j\neq i}|\rho_{ij}|. \quad (1)$$

with the symmetry constraint that $\rho_{ij} = \rho_{ji}$. For fixed ω_{ii}, (1) also forms a joint lasso regression of variable $[\rho_{12}, \rho_{13}, \cdots, \rho_{p-1,p}]^T$; for fixed ρ_{ij}, ω_{ii} can be optimized analytically. Therefore, Peng et al. (2009) proposed to optimize ρ_{ij} and ω_{ii} iteratively. Note that MB is equivalent to a simplified version of (1) in which the symmetry constraint is relaxed and ω_{ii} are set to be a uniform constant. The weights μ_i in (1) are parameters that can be manipulated to increase the flexibility of the method. If $\mu_i = \omega_{ii}$, (1) resembles a maximum pseudo likelihood estimator of the multivariate Gaussian model, which was made precise by the recent "symmetric lasso" method (Friedman et al., 2010). Peng et al. (2009) also explored various heuristics for choosing μ_i, and suggested that by taking μ_i to be the estimated degree of node i in the previous iteration, the algorithm (referred as "JSRM.dew") will prefer scale free networks. This relates to our basic task of learning scale free networks, and we return to discuss and compare this method further in the sequel.

A more systematic solution is to minimize the negative ℓ_1 penalized log-likelihood (e.g. Yuan and Lin, 2007; Banerjee et al., 2008; d'Aspremont et al., 2008; Friedman et al., 2008):

$$-\log p(x|\Omega) + \lambda\sum_{i,j}|\omega_{ij}|, \quad (2)$$

where the log-likelihood is

$$\log p(x|\Omega) = \log\det(\Omega) - \mathrm{tr}(S\Omega),$$

with $S = \frac{1}{n}XX^T$ being the empirical covariance matrix. The matrix Ω is constrained to be symmetric and positive definite during the optimization.

Optimizing (2) can be recognized as a non-differentiable convex problem. Various optimization methods have been developed to solve it. See e.g., Banerjee et al. (2008); Friedman et al. (2008); Duchi et al. (2008). One of the most efficient is the *graphical lasso* ("glasso"), which uses a blockwise coordinate descent strategy, updating each row/column of the covariance matrix by solving a modified lasso subproblem in each descent step. This method is guaranteed to maintain the property of positive definiteness during each update, so that the constraint does not need to be considered explicitly.

Across these methods there is a debate as to whether $|\omega_{ij}|$ or $|\rho_{ij}|$ should be regularized. (Note that $\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$.) In principle, JSRM can be modified to regularize $|\omega_{ij}|$, or an optimizer of the likelihood (2) such as graphical lasso could instead regularize $|\rho_{ij}|$. They are all based on the idea of ℓ_1 regularization for encouraging sparse patterns, but have minor differences in their mathematical details. Since the focus of the current work is to show how existing methods can be improved by incorporating scale-free priors, we will respect the different traditions of the original methods.

In general, all the described ℓ_1 based methods have a similar basic form, maximizing a score function with a ℓ_1 penality term

$$\mathcal{L}_{\ell_1}(\Theta) = S(X,\Theta) - \lambda||\Theta||_1, \quad (3)$$

where $||\Theta||_1 = \sum_{ij}|\theta_{ij}|$ is the ℓ_1 norm. Θ may be the precision matrix Ω or the partial correlation matrix Φ, and $S(X,\Theta)$ takes on different forms depending on the implementation details of different methods. The term $||\Theta||_1$ can be thought of as the continuous convex surrogate of the ℓ_0 norm of Θ, which equals the number of non-zero elements and thus encourages sparsity; there are also arguments that ℓ_1 is in fact the optimal convex surrogate in certain senses (see e.g., d'Aspremont et al., 2008). From a Bayesian point of view, the ℓ_1 term can be interpreted as a Laplacian prior, i.e., as assuming that the parameters θ_{ij} are distributed i.i.d. according to prior probability distribution $p(\theta_{ij}) \propto \exp(-\lambda|\theta_{ij}|)$.

3 Power Law Regularization

In scale-free network, the degree distribution of the vertices follows a power law: $p(d) \propto d^{-\alpha}$. The degree d of vertex i in a GMRF can be understood as the ℓ_0 norm of $\theta_{\neg i} = \{\theta_{ij}|j \neq i\}$. We propose to use $||\theta_{\neg i}||_1 + \epsilon_i$ as a continuous surrogate of the degree d, where $||\theta_{\neg i}||_1 = \sum_{i\neq j}|\theta_{ij}|$ is the ℓ_1 norm. The small positive term ϵ_i is used to make the surrogate positive, and hence ensure that its value in the power law distribution will be well defined. Then, we can optimize the following score function:

$$\mathcal{L}_{sf}(\Theta) = S(X,\Theta) - \alpha\sum_i\log(||\theta_{\neg i}||_1 + \epsilon_i) - \beta\sum_i|\theta_{ii}|. \quad (4)$$

The positive term ϵ_i represents the sensitivity threshold for the parameters and is used to "smooth out" the scale-free property. In practice, we take ϵ_i to have the same order as θ_{ij}. If ϵ_i is very large, i.e., $||\theta_{\neg i}||_1 \ll \epsilon_i$, then $\log(||\theta_{\neg i}||_1 + \epsilon_i) \approx \frac{1}{\epsilon_i}||\theta_{\neg i}||_1 + \log(\epsilon_i)$, and the off-diagonal regularization term becomes $\alpha \sum_{i \neq j}(\frac{1}{\epsilon_i} + \frac{1}{\epsilon_j})|\theta_{ij}|$, which is equivalent to the standard ℓ_1 regularization in (2), but with a different regularization coefficient on each element (i, j). In this sense, (4) generalizes the ℓ_1 regularization method. The diagonal elements are separated out, since we do not want to apply the power law to the diagonal. In practice, we take $\beta = 2\frac{\alpha}{\epsilon_i}$, which is the ℓ_1 regularization coefficient in the limiting case of $||\theta_{\neg i}||_1 \ll \epsilon_i$. Note that when Θ is the partial correlation, the diagonal terms are zero, and need not to be not considered.

The scale-free regularization is no longer convex as in standard ℓ_1 regularization because of the use of the log function. Intuitively, nonconvexity arises naturally in the problem of learning a scale-free network, since the locations of the hubs must be identified, creating the potential for multiple local modes.

As an alternative interpretation, the regularization in (4) can be also thought of as an approximate form of the log-normal distribution, which has been proposed as an alternative to the power law for characterizing the degree distribution of scale-free networks (Pennock et al., 2002; Mitzenmacher, 2003; Arita, 2005). A random variable z is said to be log-normal if $\log(z)$ follows a normal distribution. Formally, its log probability is

$$\log p(z) = -\log(z) - \frac{1}{2}\left(\frac{\log(z) - \mu}{\sigma}\right)^2 + \text{const.}$$

If σ is very large, the second term is small and the log-log plot of z versus $p(z)$ becomes approximately linear (Arita, 2005). Therefore, we can interpret (4) as the approximate posterior distribution after adding a log-normal prior on $(||\theta_{\neg i}||_1 + \epsilon_i)^{1/\alpha}$. Analogously to the central limit theorem, which states that the sum of many *i.i.d.* random variables approaches a Gaussian distribution, the product of many *i.i.d.* positive random variables approaches a log-normal distribution. Therefore, log-normal distributions arise quite naturally throughout the physical world, making it a reasonable prior on the parameters from a more formal Bayesian perspective.

4 Reweighted Methods

In this section, we derive an algorithm to maximize (4). We show that (4) can be solved by a sequence of ℓ_1 regularization problems, whose regularization coefficients are updated iteratively in a way that mimics the preferential attachment in scale free networks. Our method is derived as an instance of the minorize-maximize (MM) algorithm (Hunter and Lange, 2004), which is an extension of the expectation-maximization (EM) algorithm (Bilmes, 1998; Dempster et al., 1977). The MM algorithm maximizes the objective function by iteratively maximizing its minorizing functions, which are lower bounds of the objective function that are tangent to it at the current estimation point. In this way, one can always guarantee that the objective function increases monotonically and converges if it is upper bounded (Dempster et al., 1977; Wu, 1983)). More specifically, suppose Θ^n is the estimate found in the last iteration; noting that

$$\begin{aligned}&\alpha \sum_i \log(||\theta_{\neg i}||_1 + \epsilon_i) - \alpha \sum_i \log(||\theta^n_{\neg i}||_1 + \epsilon_i)\\ &\leq \alpha \sum_i \left(\frac{||\theta_{\neg i}||_1 + \epsilon_i}{||\theta^n_{\neg i}||_1 + \epsilon_i} - 1\right) = \sum_{i \neq j} \lambda_{ij}|\theta_{ij}| + \text{const.}\end{aligned}$$

where

$$\lambda_{ij} = \alpha\left(\frac{1}{||\theta^n_{\neg i}||_1 + \epsilon_i} + \frac{1}{||\theta^n_{\neg j}||_1 + \epsilon_j}\right). \tag{5}$$

then we have

$$\begin{aligned}&\mathcal{L}_{sf}(\Theta) - \mathcal{L}_{sf}(\Theta^n)\\ &\quad \geq S(X, \Theta) - \sum_{i \neq j} \lambda_{ij}|\theta_{ij}| - \beta \sum_i |\theta_{ii}| + \text{const.}\end{aligned} \tag{6}$$

The equality in (6) holds if and only if $\Theta = \Theta^n$.

Therefore, $\mathcal{L}_{sf}(\Theta)$ can be improved by iteratively maximizing the lower bound:

$$\begin{aligned}\Theta^{n+1} &= \arg\max_{\Theta} Q(\Theta|\Theta^n)\\ &= \arg\max_{\Theta}\{S(X, \Theta) - \sum_{i \neq j} \lambda_{ij}|\theta_{ij}| - \beta \sum_i |\theta_{ii}|\},\end{aligned} \tag{7}$$

which can be implemented using any of the previously discussed methods (MB, JSRM, graphical lasso, or others). It is easy to see that (6) and (7) establishes a MM algorithm, and hence the objective function increases monotonically, i.e., $\mathcal{L}_{sf}(\Theta^{n+1}) \geq \mathcal{L}_{sf}(\Theta^n)$.

This process (and the MM algorithm in general) is closely analogous to the EM algorithm, with (5) corresponding to the E-step and (7) corresponding to the M-step. The properties of EM can thus be directly applied to our algorithm. For example, the results of Wu (1983) demonstrate that $\{\Theta^n\}$ convergences to a stationary point of $\mathcal{L}_{\text{sf}}(\Theta)$, under some mild regularity conditions. We also note that it is not necessary to exactly maximize $Q(\Theta|\Theta^n)$; as with the philosophy of generalized EM (GEM) (Dempster et al., 1977), the desired properties can be maintained so long as $Q(\Theta|\Theta^n)$ is improved in each iteration. In algorithms such as graphical lasso in which the exact optimization of (7) is iterative and costly, one can update the weights more quickly, for example after one sweep of all the rows.

The weight update in (5) has an appealing intuition. It decreases the regularization strength if $||\theta_{\neg i}||_1$ is large, and hence encourages the appearance of high degree nodes (network hubs) in a rich-get-richer fashion. The process mirrors the mechanism of preferential attachment (Barabási and Albert, 1999) that has been proposed as an underlying generative model to explain power law degree distributions in scale-free networks.

5 Related Work

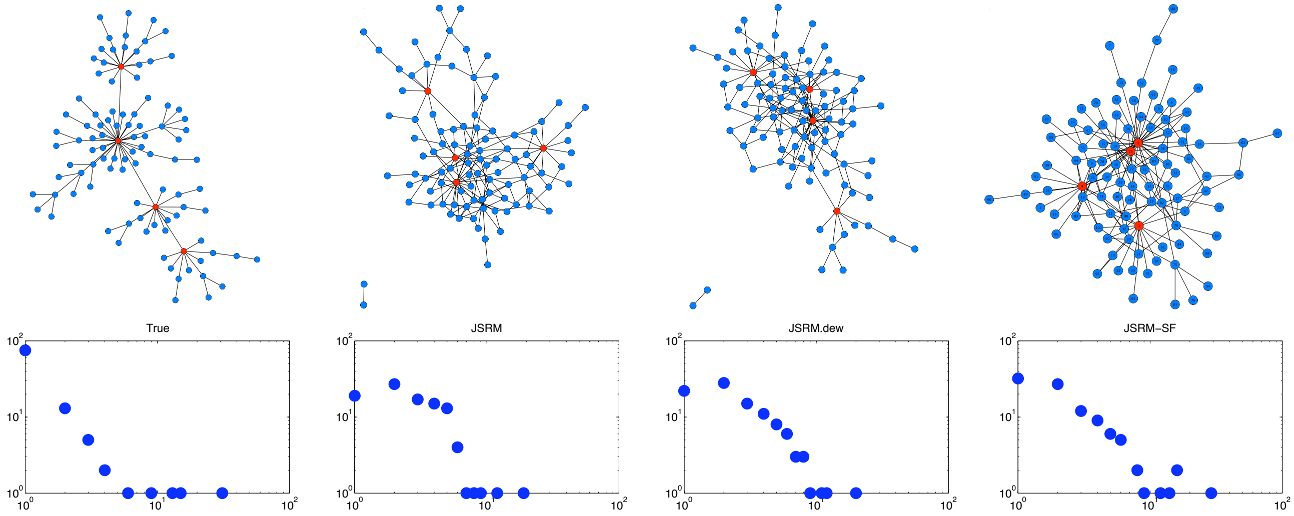
There is some existing work on learning scale free networks, but little in a systematic manner. For example, as Peng et al. (2009) suggested, JSRM can be encouraged to learn scale free network by setting the weights μ_i in (1) to be proportional to the estimated degrees in the previous iteration. This heuristic approach, however, cannot be adapted to improve other methods such as the graphical lasso. A strength of our approach is that it can be applied to improve *any* ℓ_1 based methods that have the form (3) when the underlying network is believed to be scale free. Our experimental results also suggest that our method can provide more accurate estimates than by adjusting μ_i. Another work related to constructing scale free networks is Chen et al. (2008), which is based on a simple heuristic of ranking the empirical correlation matrix.

Very recent work such as Friedman et al. (2010) and Peng et al. (2010) use a similar idea based on group lasso, where a ℓ_2 regularization on the columns of Θ is applied. These works are similar to our method in the sense that edges incident on the same node share information with each other, but induce different behaviors. Friedman et al. (2010) encourages "node sparsity", in which a node with weak correlations tends to be totally disconnected, while our method tends to reshape the network towards a power degree distribution.

An interesting connection can be drawn between our technique and the reweighted ℓ_1 minimization method for compressive sensing (Candés et al., 2008), which gives a different perspective for (5) . To illustrate, we re-write the ℓ_0 norm of some vector z as $||z||_0 = \sum_i \frac{|z_i|}{|z_i|+\epsilon_i}$, where ϵ_i is positive constant that is much smaller than the non-zero elements of z. The ℓ_1 norm $||z||_1 = \sum_i |z_i|$ has been widely used as a convex surrogate of $||z||_0$. However, it introduces large bias when the magnitude of the non-zero entries of z are very different from one another. A more accurate approximation is $||z||_{\mathrm{w}} = \sum_i \frac{|z_i|}{|z_i^0|+\epsilon_i}$, where z^0 is a previous estimate of z. Candés et al. (2008) showed that an iteratively reweighted ℓ_1 minimization is *closer* to the ground truth ℓ_0 in this sense, and presented experiments demonstrating that the reweighted algorithm leads to remarkably improved performance in applications such as sparse signal recovery. Candés et al. (2008) further noticed that the reweighted algorithm is equivalent to using the log surrogate $\sum_i \log(|z_i| + \epsilon_i)$, which also appears in our regularization. An analogous idea for lasso was explored in Shimamura et al. (2009), and was demonstrated to decrease the false positive rate when inferring a gene network. Also related in spirit is the two-pass *adaptive lasso* (Zou, 2006; Huang et al., 2006; Shimamura et al., 2007), in which z^0 is estimated using a predefined rough estimator such as shrinkage regression, but no further steps are taken to refine the weighting. Zou and Li (2008) and Fan et al. (2009) proposed a nonconcave penality function called the SCAD penality, to attenuate the bias problem of the ℓ_1 penality, and showed that the objective function can be optimized by solving a series of reweighted ℓ_1 problems. Although all these methods relate to the idea of reweighting coefficients in various ways, they are distinct from our method in that they do not explore using the weights to encourage a scale free prior of the network, but focus on attenuating the statistical bias from ℓ_1 regularization. Our preferential attachment updating is tailored to reconstruct scale free networks, which appear widely in real-world systems across many disciplines.

6 Experimental Results

We show simulation results in two types of simulated data: scale free networks simulated using a preferential attachment mechanism, and a sparse network with a few dominant hubs. Finally, we apply our method to infer a network from gene expression data, illustrating its ability to find and prefer attachments to hubs. For all the experiments in this paper, we implemented our scale free regularization on neighborhood selection ("MB"), JSRM with uniform weights ("JSRM"), and graphical lasso("glasso"). We refer to the scale free versions of MB, JSRM, and glasso as MB-SF, JSRM-SF and glasso-SF. For comparison, we also implemented the degree-reweighted version of JSRM in Peng et al. (2009) ("JSRM.dew"). For MB-SF and JSRM-SF, the variables θ_{ij} are the partial correlations $\rho_{ij} \in [-1, 1]$, and we take $\epsilon_i = 1$. For graphical lasso, $\theta_{ij} = \omega_{ij}$ are the elements of the precision matrix, we take ϵ_i equal to θ_{ii} estimated in the last iteration, to make sure ϵ_i is on the same magnitude of $||\theta_{\neg i}||_1$. For all the scale free methods, the initial



(a) True (b) JSRM (c) JSRM.dew (d) JSRM-SF

Figure 1: Top: The true scale free network and the networks estimating using JSRM (Peng et al., 2009), JSRM.dew (Peng et al., 2009), and JSRM-SF (this work). The four hubs of the original graph are shown in red in all four networks. Bottom: log-log plots of the degree distributions for each estimated network. JSRM.dew and JSRM-SF encourage scale-free degree distributions, evidenced by their linear appearance.

values of Θ are all taken to be the identity matrix so that initially $||\theta_{\neg i}||_1 = 0$. This makes the first iteration of each of the reweighting methods equivalent to its original ℓ_1 counterpart. For each method we stop after the 5^{th} reweighting iteration, although as we will show later the most significant improvement has usually been obtained even by the 2^{nd} iteration (i.e., the first reweighted iteration). The JSRM related methods are implemented in the "SPACE" R package (Peng et al., 2009).

6.1 Scale Free Network

We tested our method in random simulated scale-free networks. First, a scale-free network is simulated via the Barabási -Albert (BA) model (Barabási and Albert, 1999), which generates random scale-free networks using a preferential attachment mechanism. More specifically, the network begins with an initial, 4-node cycle. New nodes are added to the network one at a time, and each new node is connected to one of the existing nodes with a probability that is proportional to the current degree of the existing node. Formally, the probability p_i that the new node is connected to node i is $p_i = d_i / \sum_j d_j$, where d_i is the degree of i-th node. Our resulting network consisted of 100 nodes and 99 edges, with 4 hubs having degrees larger than 9; see Figure 1a.

We define $L = \eta D - G$, where G is the adjacency matrix, D is a diagonal matrix with i-th diagonal entry equal to the degree of the node i, and η is a constant larger than 1. If $\eta = 1$, L is the Laplacian matrix. We take η to be strictly larger than one (e.g., $\eta = 1.1$) to force L to be positive definite. The precision matrix Θ is then defined by $\Theta = \Lambda^{\frac{1}{2}} L \Lambda^{\frac{1}{2}}$, where Λ is the diagonal matrix of L^{-1}, which scales Θ such that the covariance matrix $\Sigma = \Theta^{-1}$ has unit diagonal, meaning that each dimension of the random vector x has equal, unit variance.

We simulated a dataset X from the Gaussian Markov model $\mathcal{N}(0, \Theta^{-1})$ with size $n = 100$. We tested MB, JSRM and glasso and their counterpart with scale free regularization on this dataset. The off-diagonal regularization coefficients α are varied to control the false positive and true positive rate for edge prediction, which yields an ROC curve. We repeat the experiment 20 times, and plot the averaged ROC curves in Figure 2. The fraction of estimated edges connecting to the hubs are shown in Figure 2 as well.

As can be seen in Figure 2 (best viewed in color), the ROC curves of the scale-free regularization methods (solid lines) are consistently above their original counterparts (dashed), and encourage a greater number of edges connecting to the hubs. Comparing the results of JSRM.dew and our JSRM-SF, it shows that the power law regularization is a more effective way of learning scale free networks.

In Figure 2c, we show the ROC curve of JSRM-SF after performing different numbers of iterations. It is evident that the greatest gain in accuracy is obtained at the 2nd iteration, i.e., with only one extra iteration

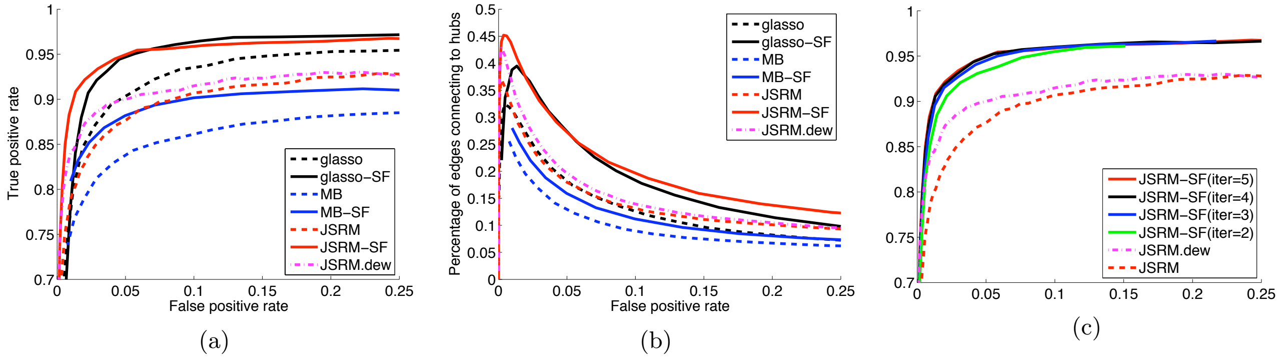


Figure 2: Experimental results on estimating the edges of a scale free network (see also Figure 1). (a) The ROC curves of each of the methods. Standard methods are shown as dashed, while their scale-free versions are shown as solid. In each case, the scale-free version significantly dominates its original; in the case of JSRM, it also dominates the scale-free JSRM.dew variant. (b) The fraction of edges connecting to the hubs vs. the false positive rate. Note that the true fraction is 0.6. Scale-free variants are more likely to find hub-connected edges "early", when few edges have been included in the graph. (c) The ROC curves of JSRM-SF after different numbers of iterations. Note that the major improvement of accuracy is obtained at the 2nd iteration (i.e., one extra iteration compared to the original ℓ_1 counterpart). This suggests that our method is not significantly more computationally expensive than the ℓ_1 counterpart.

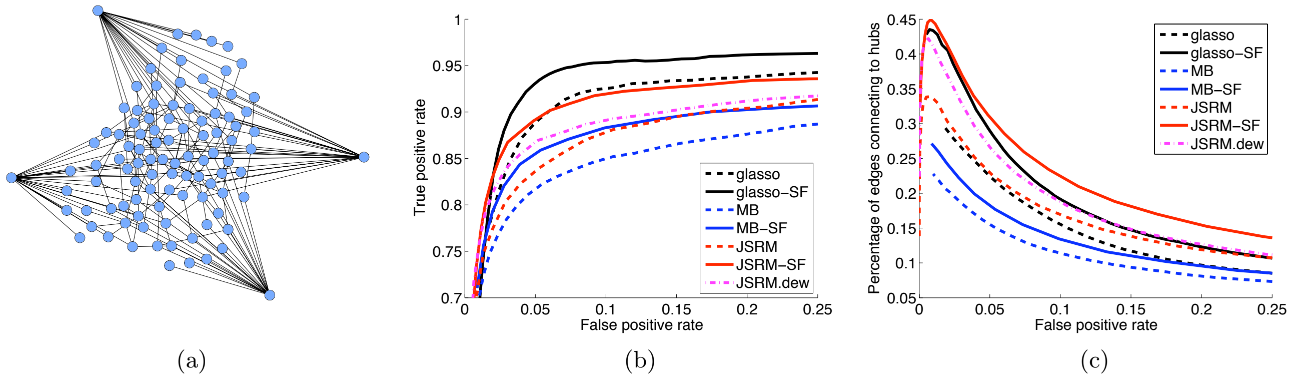
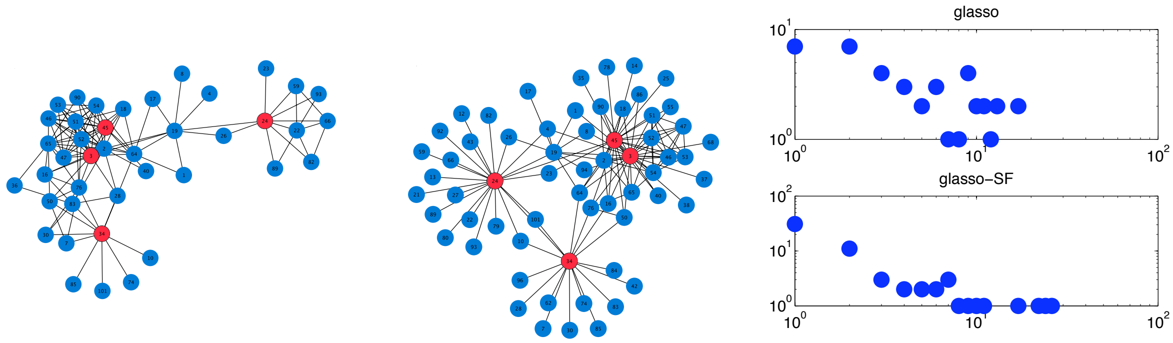


Figure 3: Experiment results on estimating the edges of a hub network. (a) The true hub network from which data are simulated. (b) The ROC curves of different methods. Standard methods are shown as dashed, while their scale-free versions are shown as solid. Again in each case, the scale-free version significantly dominates its original, as well as the scale-free JSRM.dew variant. (c) The fraction of edges connecting to the hubs vs. the false positive rate. Note that the true fraction is 0.5053. Again, we see that the scale-free variants find edges incident to the hubs earlier than their original counterparts.

compared with the original ℓ_1 method. Therefore, our scale-free iterative reweighting does not entail much additional computation over the original methods. Interestingly, similar behaviors have been found in other reweighting methods (e.g. Candés et al., 2008; Zou and Li, 2008), and it is recommended by many authors to stop after the 2nd iteration, i.e., perform only one reweighting step (Fan et al., 2009; Candés et al., 2008; Zou and Li, 2008). In case when the underlying network is not scale free, using fewer iterations may also avoid introducing too large of a bias.

We also show examples of the estimated networks found by JSRM, JSRM.dew and JSRM-SF in Figure 1. All the estimated networks have been selected to have the same number of edges (≈ 160) and thus the same false positive rate (≈ 0.025). The bottom row of Figure 1 shows log-log plots of the true degree distribution and the degree distributions of the estimated networks. Visually, the network estimated by scale free regularization most closely follows a power law; the network estimated by JSRM.dew is also similar. The results of MB vs. MB-SF and glasso vs. glasso-SF are also similar, but are not shown due to space limitations.



(a) glasso

(b) glasso-SF

(c) Degree distributions

Figure 4: Networks estimated on gene expression data. (a) The network with 120 edges estimated using glasso. (b) The network with 120 edges estimated using glasso-SF. The top 4 hubs are colored red. Both algorithms identify the same four hubs, but they are more "hub-like" in the network estiamted with glasso-SF. (c) The log-log plot of the degree distributions of the networks in (a) and (b). The network estimated using glasso-SF is noticeably more scale-free.

6.2 Hub Network

We also tested our algorithm on a sparse network with a few dominating hubs. This graph consists of four k-star subgraphs of $k = 25$ nodes, in which a hub connects to all the other nodes in its subgroup. We then add random edges between the non-hub nodes. Our network (shown in Figure 3a) results in 100 nodes and 190 edges with 4 hubs each of degree of 24. We simulated the values of the precision matrix using the same method described in Section 6.1. We draw sample data sets of size 200, and repeat the experiment 20 times to find average performance. The ROC curves of each of the different methods are shown in Figure 3b, and the fraction of edges connected to the true hubs are plotted in Figure 3c. The result strengthens the argument that our method improves accuracy compared to each of its ℓ_1 counterparts, and prefers hub structures.

6.3 Network for Gene Expression Data

We tested our algorithm on a time course profile for a set of 102 genes selected from *Saccharomyces cerevisiae* (Spellman et al., 1998; Chen et al., 2008). These microarray experiments were designed to identify yeast genes that are periodically expressed during the cell cycle. The gene expressions were collected over 18 time points, which are treated as independent samples from a GMRF in our setting. Figure 4 shows two networks with 120 edges estimated using glasso and glasso-SF, along with a log-log plot of their degree distributions (the other methods are implemented but bear similar results and are omitted for space). Visually, the network estimated using our algorithm appears closer to scale-free behavior and exhibits more clustering to the hubs. It is difficult to assess the accuracy of any of the algorithms for this problem, since the true underlying network is unknown, and existing side information is not very consistent with the data set (Chen et al., 2008; Zou and Conzen, 2005). However, we note that our methods identify the same set of highest-degree nodes (colored in red) with their ℓ_1 counterparts, but allocate more edges on the hubs (exhibiting a "preferential attachment" mechanism). This suggests that our method is consistent in the sense that it does not deviate greatly from the original methods, but imposes a slight bias toward the scale-free behavior believed to exist in the true network.

7 Conclusions and Future Directions

The study of complex networks is an active area of scientific research that examines common topological features of real-world networks. While scale-free behavior is widely acknowledged to be common, it is only one of many possible examples. Other features, such as assortativity or disassortativity among vertices, community, and hierarchical structure may also provide important information for network inference. We expect to see considerable additional cross fertilization between these two areas. There are also a number of theoretical issues of our algorithm which remain unexplored; for example, the asymptotic consistency, and the method of selecting the regularization coefficient.

Acknowledgements

This work was supported in part by NIH grants P50-GM076516 and NIAMS AR 44882 BIRT revision. The authors thank Prof. Yang Dai for sharing the microarray dataset.

References

M. Arita. Scale-freeness and biological networks. *J Biochem*, 138(1):1–4, July 2005.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:485–516, 2008.

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

A.-L. Barabási and R. Albert. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, June 2002.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc. B Stat. Meth.*, 36(2): 192–236, 1974.

J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models. Technical report, University of Berkeley, 1998.

E. J. Candés, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted L1 minimization. *J. Fourier Anal. Appl.*, 14(5), December 2008.

G. Chen, P. Larsen, E. Almasri, and Y. Dai. Rank-based edge reconstruction for scale-free genetic regulatory networks. *BMC Bioinformatics*, 9(1):75, January 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.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B*, 39(1):1–38, 1977.

J. Duchi, S. Gould, and D. Koller. Projected subgradient methods for learning sparse Gaussians. In *UAI*, 2008.

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

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1 (2):302–332, 2007.

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

J. Friedman, T. Hastie, and R. Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. 2010.

J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse highdimensional regression. Technical report, University of Iowa, 2006.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 1(58), February 2004.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009.

B. Marlin, M. Schmidt, and K. Murphy. Group sparse priors for covariance estimation. In *UAI*, 2009.

B. M. Marlin and K. P. Murphy. Sparse gaussian graphical models with unknown block structure. In *ICML*, 2009.

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

M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1(2), 2003.

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

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, 4(1), 2010.

D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, 99(8):5207–5211, April 2002.

P. Ravikumar, G. Raskutti, M. J. Wainwright, and B. Yu. High-dimensional covariance estimation by minimizing L1-penalized log-determinant. In *NIPS*, 2008.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using L1-regularized logistic regression. *Ann. Statist.*, 38(3), 2010.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. App. Gen. Mol. Biol.*, 4, 2005.

M. Schmidt, E. van den Berg, M. P. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints:a limited-memory projected quasi-newton algorithm. In *AISTATS*, volume 5, 2009.

T. Shimamura, S. Imoto, R. Yamaguchi, and S. Miyano. Weighted lasso in graphical gaussian modeling for large gene network estimation based on microarray data. *Genome Inform.*, 19, 2007.

T. Shimamura, S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano. Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3(1):41+, 2009.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, 1998.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. Royal Statist. Soc. B*, 67(1):91–108, 2005.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3), 2005.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1), 1983.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. B*, 68:49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 1, 2007.

H. Zou. The adaptive lasso and its oracle properties. *JASA*, 101:1418–1429, December 2006.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36(4), 2008.

M. Zou and S. D. Conzen. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, January 2005.