High-dimensional Sparse Inverse Covariance Estimation using Greedy Methods

Christopher C. Johnson
CS, UT Austin
cjohnson@cs.utexas.edu

Ali Jalali
ECE, UT Austin
alij@mail.utexas.edu

Pradeep Ravikumar
CS, UT Austin
pradeepr@cs.utexas.edu

Abstract

In this paper we consider the task of estimating the non-zero pattern of the sparse inverse covariance matrix of a zero-mean Gaussian random vector from a set of iid samples. Note that this is also equivalent to recovering the underlying graph structure of a sparse Gaussian Markov Random Field (GMRF). We present two novel greedy approaches to solving this problem. The first estimates the non-zero covariates of the overall inverse covariance matrix using a series of global forward and backward greedy steps. The second estimates the neighborhood of each node in the graph separately, again using greedy forward and backward steps, and combines the intermediate neighborhoods to form an overall estimate. The principal contribution of this paper is a rigorous analysis of the sparsistency of these two greedy procedures, that is, their consistency in recovering the sparsity pattern of the inverse covariance matrix. Surprisingly, we show that both the local and global greedy methods learn the full structure of the model with high probability given just $O(d\log(p))$ samples, which is a *significant* improvement over state of the art ℓ_1-regularized Gaussian MLE (Graphical Lasso) that requires $O(d^2\log(p))$ samples. Moreover, the restricted eigenvalue and smoothness conditions imposed by our greedy methods are much weaker than the strong irrepresentable conditions required by the ℓ_1-regularization based methods. We corroborate our results with extensive simulations and examples, comparing our local and global greedy methods to the ℓ_1-regularized Gaussian MLE as well as the nodewise ℓ_1-regularized linear regression (Neighborhood Lasso).

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

1 Introduction

High-dimensional Covariance Estimation. Increasingly, modern statistical problems across varied fields of science and engineering involve a large number of variables. Estimation of such high-dimensional models has been the focus of considerable recent research, and it is now well understood that consistent estimation is possible when some low-dimensional structure is imposed on the model space. In this paper, we consider the specific high-dimensional problem of recovering the covariance matrix of a zero-mean Gaussian random vector, under the low-dimensional structural constraint of *sparsity* of the inverse covariance, or concentration matrix. When the random vector is multivariate Gaussian, the set of non-zero entries in the concentration matrix correspond to the set of edges in an associated Gaussian Markov random field (GMRF). In this setting, imposing sparsity on the entries of the concentration matrix can be interpreted as requiring that the graph underlying the GMRF have relatively few edges.

State of the art: ℓ_1 regularized Gaussian MLE. For this task of sparse GMRF estimation, a line of recent papers [3, 5, 15] have proposed an estimator that minimizes the Gaussian negative log-likelihood regularized by the ℓ_1 norm of the entries (or the off-diagonal entries) of the concentration matrix. The resulting optimization problem is a log-determinant program, which can be solved in polynomial time with interior point methods [1], or by co-ordinate descent algorithms [3, 5]. Rothman et al. [11], Ravikumar et al. [10] have also shown strong statistical guarantees for this estimator: both in ℓ_2 operator norm error bounds, and recovery of the underlying graph structure.

Recent resurgence of greedy methods. A related line of recent work on learning sparse models has focused on "stagewise" greedy algorithms. These perform simple forward steps (adding parameters greedily), and possibly also backward steps (removing parameters greedily), and yet provide strong statistical guarantees for the estimate after a finite number of greedy steps. Indeed, such greedy algorithms have appeared in various guises in multiple communities: in machine learning as boosting [4], in function approximation [13], and in signal processing as basis pursuit [2]. In the context of statistical model estimation, Zhang [17] analyzed the forward greedy algorithm for the case of sparse linear regression; and showed that the forward greedy algorithm is sparsistent (consistent for model selection recovery) under the same "irrepresentable" condition as that required for "sparsistency" of the Lasso. Zhang [16] analyzes a more general greedy algorithm for sparse linear regression that performs forward and backward steps, and showed that it is sparsistent under a weaker restricted eigenvalue condition. Jalali et al. [7] extend the sparsistency analysis of [16] to general non-linear models, and again show that strong sparsistency guarantees hold for these algorithms.

Our Approaches. Motivated by these recent results, we apply the forward-backward greedy algorithm studied in [16, 7] to the task of learning the graph structure of a Gaussian Markov random field given iid samples. We propose two algorithms: one that applies the greedy algorithm to the overall Gaussian log-likelihood loss, and the other that is based on greedy neighborhood estimation. For this second method, we follow [8, 9], and estimate the neighborhood of each node by applying the greedy algorithm to the local node conditional log-likelihood loss (which reduces to the least squares loss), and then show that each neighborhood is recovered with very high probability, so that by an elementary union bound, the entire graph structure is recovered with high probability. A principal contribution of this paper is a rigorous analysis of these algorithms, where we report sufficient conditions for recovery of the underlying graph structure. We also corroborate our analysis with extensive simulations.

Our analysis shows that for a Gaussian random vector $X = (X_1, X_2, \ldots, X_p)$ with p variables, both the global and local greedy algorithms only require $n = O(d \log(p))$ samples for sparsistent graph recovery. Note that this is a significant improvement over the ℓ_1 regularized Gaussian MLE [15] which has been shown to require $O(d^2 \log(p))$ samples [10]. Moreover, we show that the local and global greedy algorithms require a very weak restricted eigenvalue and restricted smoothness condition on the true inverse covariance matrix (with the local greedy imposing a marginally weaker condition that the global greedy algorithm). This is in contrast to the ℓ_1 regularized Gaussian MLE which imposes a very stringent edge-based irrepresentable condition [10]. In Section 5, we explicitly compare these different conditions imposed by the various methods for some simple GMRFs, and quantitatively show that the conditions imposed by the local and global greedy methods require much weaker conditions on the covariance entries. Thus, both theoretically and via simulations, we show that the set of methods proposed in the paper are the *state of the art* in recovering the graph structure of a GMRF from iid samples: both in the number of samples required, and the weakness of the sufficient conditions imposed upon the model.

2 Problem Setup

2.1 Gaussian graphical models

Let $X = (X_1, X_2, \ldots, X_p)$ be a zero-mean Gaussian random vector. Its density is parameterized by its inverse covariance or *concentration matrix* $\Theta^* = (\Sigma^*)^{-1} \succ 0$, and can be written as

$$f(x_1, \ldots, x_p; \Theta^*) = \frac{\exp\big\{ -\frac{1}{2} x^T \Theta^* x \big\}}{\sqrt{(2\pi)^p \det(\Theta^*)^{-1}}}. \qquad (1)$$

We can associate an undirected graph structure $G = (V, E)$ with this distribution, with the vertex set $V = \{1, 2, \ldots, p\}$ corresponding to the variables $(X_1, \ldots, X_p)$, and with edge set such that $(i, j) \notin E$ if $\Theta^*_{ij} = 0$.

We are interested in the problem of recovering this underlying graph structure, which corresponds to determining which off-diagonal entries of Θ^* are non-zero—that is, the set

$$E(\Theta^*) \;\; := \;\; \{i, j \in V \mid i \neq j, \Theta^*_{ij} \neq 0\}. \qquad (2)$$

Given n samples, we define the *sample covariance matrix*

$$\widehat{\Sigma}^n \;\; := \;\; \frac{1}{n} \sum_{k=1}^{n} X^{(k)} (X^{(k)})^T. \qquad (3)$$

In the sequel, we occasionally drop the superscript n, and simply write $\widehat{\Sigma}$ for the sample covariance.

With a slight abuse of notation, we define the *sparsity index* $s := |E(\Theta^*)|$ as the total number of non-zero elements in off-diagonal positions of Θ^*; equivalently, this corresponds to twice the number of edges in the case of a Gaussian graphical model. We also define the *maximum degree or row cardinality*

$$d \;\; := \;\; \max_{i=1,\ldots,p} \Big| \big\{ j \in V \mid \Theta^*_{ij} \neq 0 \big\} \Big|, \qquad (4)$$

corresponding to the maximum number of non-zeros in any row of Θ^*; this corresponds to the maximum degree in the graph of the underlying Gaussian graphical model. Note that we have included the diagonal entry Θ^*_{ii} in the degree count, corresponding to a self-loop at each vertex.

2.2 State of the art: ℓ_1 regularization

Define the *off-diagonal* ℓ_1 *regularizer*

$$\|\Theta\|_{1,\text{off}} \quad := \quad \sum_{i \neq j} |\Theta_{ij}|, \tag{5}$$

where the sum ranges over all $i, j = 1, \ldots, p$ with $i \neq j$. Given some regularization constant $\lambda_n > 0$, we consider estimating Θ^* by solving the following ℓ_1-*regularized log-determinant program*:

$$\widehat{\Theta} := \arg \min_{\Theta \in \mathcal{S}^p_{++}} \left\{ \langle\!\langle \Theta,\, \widehat{\Sigma}^n \rangle\!\rangle - \log \det(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \right\}, \tag{6}$$

which returns a symmetric positive definite matrix $\widehat{\Theta}$.

Note that this corresponds to the ℓ_1 regularized Gaussian MLE when the underlying distribution is Gaussian.

2.3 Forward Backward Greedy

[16, 7] consider a simple forward-backward greedy algorithm for model estimation that begins with an empty set of active variables and gradually adds (and removes) variables to the active set. This algorithm has two basic steps: the forward step and the backward step. In the forward step, the algorithm finds the *best* next candidate and adds it to the active set as long as it improves the loss function at least by $\epsilon_{\mathcal{S}}$, otherwise the stopping criterion is met and the algorithm terminates. Then, in the backward step, the algorithm checks the *influence* of all variables in the presence of the newly added variable. If one or more of the previously added variables do not contribute at least $\nu\epsilon_{\mathcal{S}}$ to the loss function, then the algorithm removes them from the active set. This procedure ensures that at each round, the loss function is improved by at least $(1-\nu)\epsilon_{\mathcal{S}}$ and hence it terminates within a finite number of steps.

In the sequel, we will apply this greedy methodology to Gaussian graphical models, to obtain two methods: (a) Greedy Gaussian MLE, which applies the greedy algorithm to the Gaussian negative log-likelihood loss, and (b) Greedy Neighborhood Estimation, which applies the greedy algorithm to the local node-conditional negative log-likelihood loss.

Algorithm 1 Global greedy forward-backward algorithm for Gaussian covariance estimation

Input: $\widehat{\Sigma}^n$, Stopping Threshold ϵ_S, Backward Step Factor $\nu \in (0, 1)$
Output: Inverse Covariance Estimation $\widehat{\Theta}$

Initialize $\widehat{\Theta}^{(0)} \leftarrow \mathbb{I}$, $\widehat{S}^{(0)} \leftarrow \emptyset$, and $k \leftarrow 1$

while true **do** {*Forward Step*}
  $((i_*, j_*), \alpha_*) \longleftarrow \arg \min_{(i,j) \in \left(\widehat{S}^{(k-1)}\right)^c; \alpha} \mathcal{L}\left(\widehat{\Theta}^{(k-1)} + \alpha(e_{ij} + e_{ji})\right)$
  $\widehat{S}^{(k)} \longleftarrow \widehat{S}^{(k-1)} \cup \{(i_*, j_*)\}$
  $\delta_f^{(k)} \longleftarrow \mathcal{L}(\widehat{\Theta}^{(k-1)}) - \mathcal{L}\left(\widehat{\Theta}^{(k-1)} + \alpha_*(e_{i_* j_*} + e_{j_* i_*})\right)$
  if $\delta_f^{(k)} \leq \epsilon_S$ **then**
    break
  end if
  $\widehat{\Theta}^{(k)} \longleftarrow \arg\min_{\Theta} \mathcal{L}\left(\Theta_{\widehat{S}^{(k)}}\right)$
  $k \longleftarrow k + 1$
  while true **do** {*Backward Step*}
    $(i^*, j^*) \longleftarrow \arg \min_{j \in \widehat{S}^{(k-1)}} \mathcal{L}\left(\widehat{\Theta}^{(k-1)} - \widehat{\Theta}^{(k-1)}_{ij}(e_{ij} + e_{ji})\right)$
    if $\mathcal{L}\left(\widehat{\Theta}^{(k-1)} - \widehat{\Theta}^{(k-1)}_{i^* j^*}(e_{i^* j^*} + e_{j^* i^*})\right) - \mathcal{L}\left(\widehat{\Theta}^{(k-1)}\right) > \nu\delta_f^{(k)}$ **then**
      break
    end if
    $\widehat{S}^{(k-1)} \longleftarrow \widehat{S}^{(k)} - \{(i^*, j^*)\}$
    $\widehat{\Theta}^{(k-1)} \longleftarrow \arg\min_{\theta} \mathcal{L}\left(\Theta_{\widehat{S}^{(k-1)}}\right)$
    $k \longleftarrow k - 1$
  end while
end while

3 Greedy Gaussian MLE

In Algorithm 1, we describe the greedy algorithm of [16, 7] as applied to the Gaussian log-likelihood loss,

$$\mathcal{L}(\Theta) := \langle\!\langle \Theta,\, \widehat{\Sigma}^n \rangle\!\rangle - \log \det(\Theta).$$

Assumption:
Let $\rho \geq 1$ be a constant and $\Delta \in \mathbb{R}^{p \times p}$ be a symmetric matrix that is *sparse* with at most ηd nonzero entries per row (and column) for some $\eta \geq 2 + 4\rho^2(\sqrt{(\rho^2 - \rho)/d} + \sqrt{2})^2$. We require that population covariance matrix $\Sigma^* = \mathbb{E}\left[XX^T\right]$ satisfy the restricted eigenvalue property, i.e., for some positive constants $C_{\min}$, we have

$$C_{\min} \|\Delta\|_F \leq \langle\!\langle \Sigma^*,\, \Delta \rangle\!\rangle \leq \rho C_{\min} \|\Delta\|_F,$$

where, $\|\cdot\|_F$ denotes the Frobenius norm.

Lemma 1. *Suppose Σ^* satisfies the assumption in 7. Then, with probability at least $1 - c_1 \exp(-c_2 n)$ for arbitrary small constant $\alpha > 0$, we have that for any symmetric matrix Δ with ηd non-zero entries per row (and column),*

$$(1-\alpha)C_{\min}\|\Delta\|_F \le \langle\langle \widehat{\Sigma}^n,\ \Delta\rangle\rangle \le (1+\alpha)\rho C_{\min}\|\Delta\|_F,$$

provided that $n \ge K\, d\, \log(p)$ for some positive constant K, c_1 and c_2.

Proof. The proof follows from Lemma 9 (Appendix K) in [14]. □

Using Taylor series expansion, we can write

$$\mathcal{L}(\Theta+\Delta) = \mathcal{L}(\Theta) + \langle\langle \Delta,\ \widehat{\Sigma}^n\rangle\rangle - \langle\langle \Theta^{-1},\ \Delta\rangle\rangle + \underbrace{\sum_{i=2}^{\infty} \frac{(-1)^i}{i} \langle\langle\ (\Theta^{-1}\Delta)^{i-1}\,\Theta^{-1},\ \Delta\rangle\rangle}_{R_\Delta}.$$

In order to establish the restricted strong convexity/smoothness required by [7], we need to lower/upper bound R_Δ. Notice that in the proof of [7], the required Δ is the difference between the target variable Θ^* and the k^{th} step estimation $\widehat{\Theta}^{(k)}$. Since the algorithm is guaranteed to converge, $\Delta = \Theta^* - \widehat{\Theta}^{(k)}$ is always bounded. Thus, without loss of generality, we assume that $\|\Delta\|_F \le 1$. Notice that we can scale $\|\Delta\|_F$ and similar type of result holds. The next lemma provides the required upper/lower bound.

Lemma 2. *Suppose Σ^* satisfies the assumption in 7. Then with probability at least $1-c_1\exp(-c_2 n)$, we have that for any symmetric matrix Δ with ηd non-zero entries per row (and column), and with $\|\Delta\|_F \le 1$,*

$$\frac{1}{4}C_{\min}^2\|\Delta\|_F^2 \le R_\Delta \le \frac{1}{2}\rho^2 C_{\min}^2\|\Delta\|_F^2.$$

Proof. Denote $\gamma = \langle\langle \Theta^{-1},\ \Delta\rangle\rangle$. We have

$$R_\Delta = \sum_{i=2}^{\infty}\frac{(-1)^i}{i}\gamma^i = \gamma - \log(1+\gamma).$$

Under our assumption, $C_{\min}\|\Delta\|_F \le \gamma \le \rho C_{\min}\|\Delta\|_F$ and the function $\gamma - \log(1+\gamma)$ is an increasing function in γ. Moreover, for the range of γ, we have $\gamma - \log(1+\gamma) \ge \frac{1}{4}\gamma^2$ because they both vanish at zero and the derivative of LHS is larger than the derivative of LHS. Hence, we have

$$\begin{aligned}\frac{1}{4}C_{\min}^2\|\Delta\|_F^2 &\le C_{\min}\|\Delta\|_F - \log(1+C_{\min}\|\Delta\|_F)\\ &\le \gamma - \log(1+\gamma) = R_\Delta\\ &\le \rho C_{\min}\|\Delta\|_F - \log(1+\rho C_{\min}\|\Delta\|_F)\\ &\le \frac{1}{2}\rho^2 C_{\min}^2\|\Delta\|_F^2.\end{aligned}$$

The last inequality follows from $\gamma - \log(1+\gamma) \le \frac{1}{2}\gamma^2$ (since they are equal at zero and the derivative of RHS is always larger above zero). Hence, the result follows. □

Let $\nabla^{(n)} := \|\widehat{\Sigma}^n - (\Theta^*)^{-1}\|_\infty$. By first order condition on the optimality of Θ^*, it is clear that $\lim_{n\to\infty}\nabla^{(n)} = 0$. The following lemma provides an upper bound on $\nabla^{(n)}$.

Lemma 3. *Given the sample complexity $n \ge K\,\log(p)$ for some constant K, we have*

$$\nabla^{(n)} \le c\sqrt{\frac{\log(p)}{n}},$$

with probability at least $1-c_1\exp(-c_2 n)$ for some positive constants c, c_1 and c_2.

Proof. The proof follows from Lemma 1 in [10]. □

This entails that the restricted strong convexity and smoothness (i.e., the required assumptions of the general result in [7]) are satisfied. Now, we can specialize the results in [7] to obtain the following theorem:

Theorem 1 (Global Greedy Sparsistency)**.** *Under the assumption above, suppose we run Algorithm 1 with stopping threshold $\epsilon_{\mathcal{S}} \ge (2c\eta/\rho^2)d\log(p)/n$, where, d is the maximum node degree in the graphical model, and the true parameters Θ^* satisfy $\min_{t\in\mathcal{S}^*}|\Theta^*| \ge \sqrt{8\epsilon_{\mathcal{S}}/\rho^2}$, and further that number of samples scales as*

$$n > K\, d\, \log(p)$$

for some constant K. Then, with probability at least $1 - c_1\exp(-c_2 n)$, we have

(a) **No False Exclusions:** $E^* - \widehat{E} = \emptyset$.

(b) **No False Inclusions:** $\widehat{E} - E^* = \emptyset$.

4 Greedy Neighborhood Estimation

Denote by $\mathcal{N}^*(r)$ the set of neighbors of a vertex $r \in V$, so that $\mathcal{N}^*(r) = \{t : (r,t) \in E^*\}$. Then the graphical model selection problem is equivalent to that of estimating the neighborhoods $\hat{\mathcal{N}}_n(r) \subset V$, so that $\mathbb{P}[\hat{\mathcal{N}}_n(r) = \mathcal{N}^*(r); \forall r \in V] \to 1$ as $n \to \infty$.

For any pair of random variables X_r and X_t, the parameter Θ_{rt} fully characterizes whether there is an edge between them, and can be estimated via its conditional likelihood. In particular, defining $\Theta_r := \{\Theta_{rt}\}_{t\neq r}$, our goal is to use the conditional likelihood of X_r conditioned on $X_{V\setminus r}$ to estimate the *support* of Θ_r and hence its neighborhood $\mathcal{N}(r)$. This conditional

distribution of X_r conditioned on $X_{V\setminus r}$ generated by (1) is given by (considering $\Theta^{-1} = \Sigma$)

$$X_r | X_{V\setminus r} \sim \mathcal{N}\left(-\Theta_{r\setminus r}^{-1}\Theta_{\setminus r\setminus r}\, X_{V\setminus r},\ \Theta_{rr}^{-1} - \Theta_{r\setminus r}^{-1}\Theta_{\setminus r\setminus r}\Theta_{\setminus r\, r}^{-1}\right).$$

However, note that we do not need to estimate the variance of this conditional distribution in order to obtain the support of $\Theta_r = \Theta_{r\setminus r}$. In particular, the solution to the following least squares loss

$$\Gamma_r^* = \arg\min_{\Gamma_r} \mathbb{E}[(X_r - \sum_{t\neq r} \Gamma_{rt} X_t)^2],$$

would satisfy $\mathrm{supp}(\Gamma_r^*) = \mathrm{supp}(\Theta_r^*)$.

Given the n samples $X^{(1)}, \ldots, X^{(n)}$, we thus use the sample-based linear loss

$$\mathcal{L}(\Gamma_r) = \frac{1}{2n}\sum_{i=1}^{n}\left(X_r^{(i)} - \sum_{t\neq r}\Gamma_{rt}X_t^{(i)}\right)^2. \qquad (7)$$

Adapting the greedy algorithm from the previous section to this linear loss at each node thus yields Algorithm 2.

Assumption:
Let $\rho \geq 1$ be a constant and $\Delta \in \mathbb{R}^{p-1}$ be an arbitrary ηd-sparse vector, where, $\eta \geq 2 + 4\rho^2(\sqrt{(\rho^2-\rho)/d} + \sqrt{2})^2$. We require the marginal population Fisher information matrix $\Sigma_{\setminus r}^* = \mathbb{E}\left[X_{\setminus r}X_{\setminus r}^T\right]$ satisfy the restricted eigenvalue property, i.e., for some positive constants $C_{\min}$, we have

$$C_{\min}\|\Delta\|_F \leq \|\Sigma_{\setminus r}^*\Delta\|_F \leq \rho C_{\min}\|\Delta\|_F.$$

Lemma 4. *Under assumption above, and for some arbitrary small constant $\alpha > 0$, the marginal sample Fisher information matrix $\widehat{\Sigma}_{\setminus r}^n = \frac{1}{n}\sum_{i=1}^n X_{\setminus r}^{(i)}X_{\setminus r}^{(i)T}$, with probability at least $1 - c_1\exp(-c_2 n)$, satisfies the condition that for any symmetric matrix Δ with ηd non-zero entries per row (and column),*

$$(1-\alpha)C_{\min}\|\Delta\|_F \leq \|\widehat{\Sigma}_{\setminus r}^n\Delta\|_F \leq (1+\alpha)\rho C_{\min}\|\Delta\|_F,$$

provided that $n \geq K\, d\log(p)$ for some positive constant K, c_1 and c_2.

Proof. The proof follows from Lemma 9 (Appendix K) in [14]. □

Let $\nabla_r^{(n)} := \max_t \left|\frac{1}{n}\sum_{i=1}^n X_t^{(i)}\left(X_r^{(i)} - \sum_{t\neq r}\Gamma_{rt}^* X_t^{(i)}\right)\right|$. By first order condition on the optimality of Γ_{rt}^*, it is clear that $\lim_{n\to\infty}\nabla_r^{(n)} = 0$. The following lemma provides an upper bound on $\nabla_r^{(n)}$.

Algorithm 2 Greedy forward-backward algorithm for marginal Gaussian covariance estimation

Input: Data Vectors $X^{(1)}, \ldots, X^{(n)}$, Stopping Threshold ϵ_S, Backward Step Factor $\nu \in (0,1)$
Output: Marginal Vector $\widehat{\Gamma}_r$

Initialize $\widehat{\Gamma}_r^{(0)} \leftarrow \mathbf{0}$, $\widehat{S}^{(0)} \leftarrow \emptyset$, and $k \leftarrow 1$
while true **do** {*Forward Step*}
  $(t_*, \alpha_*) \longleftarrow \arg\min_{t\in(\widehat{S}^{(k-1)})^c;\,\alpha} \mathcal{L}\left(\widehat{\Gamma}_r^{(k-1)} + \alpha e_t\right)$
  $\widehat{S}^{(k)} \longleftarrow \widehat{S}^{(k-1)} \cup \{t_*\}$
  $\delta_f^{(k)} \longleftarrow \mathcal{L}(\widehat{\Gamma}_r^{(k-1)}) - \mathcal{L}(\widehat{\Gamma}_r^{(k-1)} + \alpha_* e_{t_*})$
  if $\delta_f^{(k)} \leq \epsilon_S$ **then**
    break
  end if

  $\widehat{\Gamma}_r^{(k)} \longleftarrow \arg\min_{\Gamma_r} \mathcal{L}\big((\Gamma_r)_{\widehat{S}^{(k)}}\big)$
  $k \longleftarrow k+1$

  while true **do** {*Backward Step*}
    $t^* \longleftarrow \arg\min_{t\in\widehat{S}^{(k-1)}} \mathcal{L}(\widehat{\Gamma}_r^{(k-1)} - \widehat{\Gamma}_{rt}^{(k-1)} e_t)$
    if $\mathcal{L}\big(\widehat{\Gamma}_r^{(k-1)} - \widehat{\Gamma}_{rt^*}^{(k-1)} e_{t^*}\big) - \mathcal{L}\big(\widehat{\Gamma}_r^{(k-1)}\big) > \nu\delta_f^{(k)}$ **then**
      break
    end if

    $\widehat{S}^{(k-1)} \longleftarrow \widehat{S}^{(k)} - \{t^*\}$
    $\widehat{\Gamma}_r^{(k-1)} \longleftarrow \arg\min_{\Gamma_r} \mathcal{L}\big((\Gamma_r)_{\widehat{S}^{(k-1)}}\big)$
    $k \longleftarrow k-1$
  end while
end while

Lemma 5. *Given the sample complexity $n \geq K\,\log(p)$ for some constant K, we have*

$$\nabla_r^{(n)} \leq c\sqrt{\frac{\log(p)}{n}},$$

with probability at least $1 - c_1\exp(-c_2 n)$ for some positive constants c, c_1 and c_2.

Proof. The proof follows from Lemma 5 in [14]. □

This entails that the restricted strong convexity and smoothness (i.e., the required assumptions of the general result in [7]) are satisfied with constants $C_{\min}$ and $\rho C_{\min}$, respectively; because, the third and higher order derivatives are zero. Now, we can then specialize the results in [7] to obtain the following theorem:

Theorem 2 (Neighborhood Greedy Sparsistency)**.** *Under the assumption above, suppose we run Algorithm 2 with stopping threshold $\epsilon_{\mathcal{S}} \geq (8c\rho\eta/C_{\min})d\log(p)/n$, where, d is the maximum*

node degree in the graphical model, and the true parameters Γ^*_r *satisfy* $\min_{t\in\mathcal{N}^*(r)} |\Gamma^*_{rt}| \geq \sqrt{32\rho\epsilon_{\mathcal{S}}/C_{\min}}$, *and further that number of samples scales as*

$$n > K\, d\, \log(p)$$

for some constant K. *Then, with probability at least* $1 - c_1 \exp(-c_2 n)$, *we have*

(a) **No False Exclusions:** $E^*_r - \widehat{E}_r = \emptyset$.

(b) **No False Inclusions:** $\widehat{E}_r - E^*_r = \emptyset$.

5 Comparisons to Related Methods

In this section, we compare our global and local greedy methods to the ℓ_1-regularized Gaussian MLE, analyzed in [10], and to ℓ_1-regularization (Lasso) based neighborhood selection, analyzed in [8, 14].

5.1 Sample Complexity

Our greedy algorithm requires $\mathcal{O}(d\,\log(p))$ samples to recover the exact structure of the graph for both the global and local neighborhood based methods. In contrast, the ℓ_1-regularized Gaussian MLE [10] requires $\mathcal{O}(d^2\,\log(p))$ samples to guarantee structure recovery with high probability. The linear neighborhood selection with ℓ_1-regularization [8] requires $\mathcal{O}(d\,\log(p))$ samples to guarantee sparsistency, similar to our greedy algorithms.

5.2 Minimum Non-Zero Values

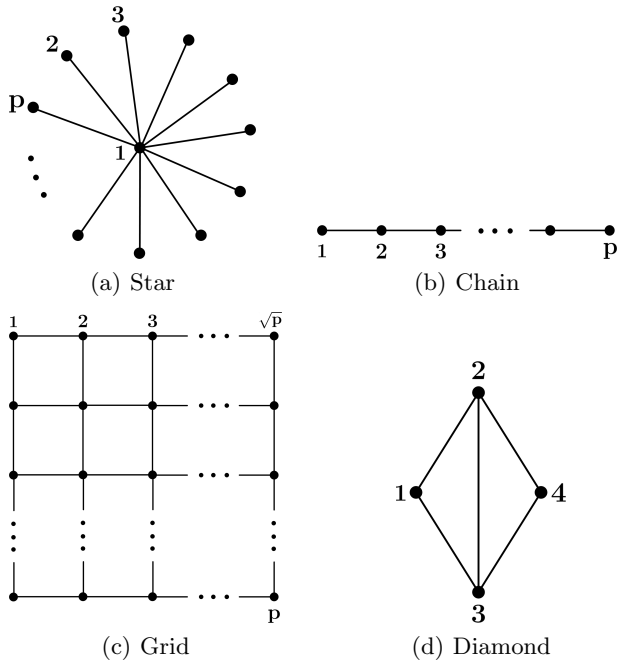
The ℓ_1-regularized Gaussian MLE imposes the model condition that the minimum non-zero entry of $\Sigma^{*\,-1}$ satisfy $\Sigma^{*\,-1}_{\min} = \mathcal{O}(1/d)$. Our greedy algorithms allow for a broader range of minimum non-zero values $\Sigma^{*\,-1}_{\min} = \mathcal{O}(1/\sqrt{d})$. The linear neighborhood selection with ℓ_1-regularization again matches our greedy algorithms and only requires that $\Sigma^{*\,-1}_{\min} = \mathcal{O}(1/\sqrt{d})$.

5.3 Parameter Restrictions

We now compare the irrepresentable and restricted eigenvalue and smoothness conditions imposed on the model parameters by the different methods.

5.3.1 Star Graphs

Consider a star graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ with p nodes in Fig 1(a), where the center node is labeled 1 and the other nodes are labeled from 2 to p. Following [10], consider the following covariance matrix Σ^* parameterized by the correlation parameter $\tau \in [-1,1]$: the diagonal entries are set to $\Sigma^*_{ii} = 1$, for all $i \in V$; the entries corresponding to edges are set to $\Sigma^*_{ij} = \tau$ for $(i,j) \in E$; while the non-edge entries are set as $\Sigma^*_{ij} = \tau^2$ for $(i,j) \notin E$. It is easy to check that Σ^* induces the desired star graph. With this setup, the irrepresentable condition imposed by the ℓ_1-regularized Gaussian MLE [10] entails that $|\tau|(|\tau|+2) < 1$ or equivalently $\tau \in (-0.4142, 0.4142)$ to guarantee sparsistency. However, our greedy algorithms allow for $\tau \in (-1,1)$ (since $C_{\min} = 1-\tau^2$). Under the same setup, the linear neighborhood selection with ℓ_1-regularization [8] requires $\tau \in (-1,1)$ to guarantee the success.

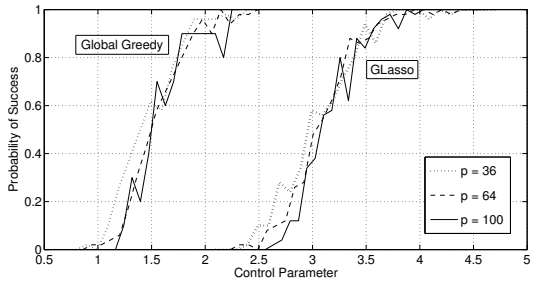


(a) Star (b) Chain (c) Grid (d) Diamond

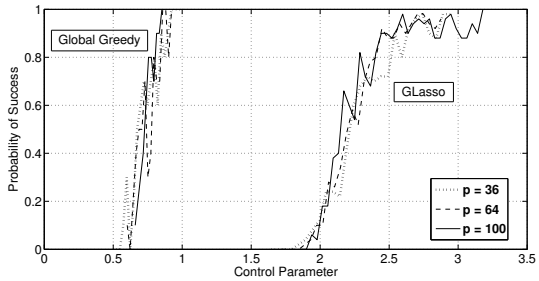
Figure 1: Generic Graph Schematics

5.3.2 Chain Graphs

Consider a chain (line) graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ on p nodes as shown in Fig 1(b). Again, consider a population covariance matrix Σ^* parameterized by the correlation parameter $\tau \in [-1,1]$: set $\Sigma^*_{ij} = \tau^{|i-j|}$. Thus, this matrix assumes a correlation factor of τ^k between two nodes that are k hops away from each other. It is easy to check that Σ^* induces the desired chain graph. With this setup, the ℓ_1-regularized Gaussian MLE [10] requires $|\tau|^{p-2}\,((p-2)|\tau| + p - 1) < 1$. It is hard to evaluate bounds on τ in general, but for the case $p = 4$ we have $\tau \in (-0.6, 0.6)$; for the case $p = 10$ we have $\tau \in (-0.75, 0.75)$ and for the case $p = 100$ we have $\tau \in (-0.95, 0.95)$. Our greedy algorithms on the other hand allow for $\tau \in (-1,1)$ (since $C_{\min} = (1-\tau^2) f_p(\tau)$ for some function $f_p(\tau)$ that depends on p and satisfies $f_p(\tau) > C_p$ for all τ and some constant C_p depending only on p). Under the same setup, the linear neighborhood selection with ℓ_1-regualrization [8] only imposes

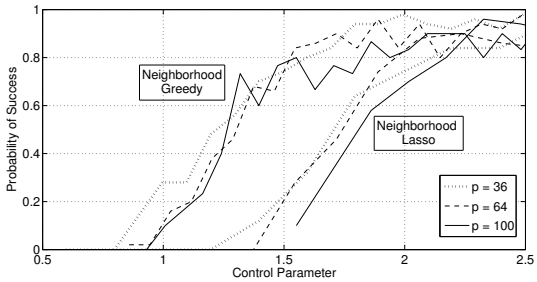


(a) Chain (Line Graph)

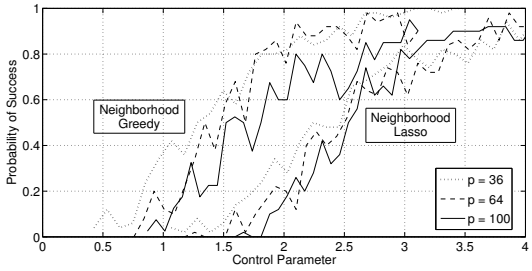


(b) 4-Nearest Neighbor (Grid Graph)

Fig 2: Plots of success probability $\mathbb{P}[\widehat{S} = S^*]$ versus the control parameter $\beta(n, p, d) = n/[70d\log(p)]$ for (a) chain ($d = 2$) and (b) 4-nearest neighbor grid ($d = 4$) using both Algorithm 1 and ℓ_1-regularized Gaussian MLE (Graphical Lasso). As our theorem suggests and these figures show, the Global Greedy algorithm requires less samples to recover the exact structure of the graphical model.



(a) Chain (Line Graph)



(b) Star Graph

Fig 3: Plots of success probability $\mathbb{P}[\widehat{\mathcal{N}}_{\pm}(r) = \mathcal{N}^*(r), \forall r \in V]$ versus the control parameter $\beta(n, p, d) = n/[70d\log(p)]$ for (a) chain ($d = 2$) and $\beta(n, p, d) = n/[200\log(dp)]$ for (b) star graph ($d = 0.1p$) using both Algorithm 2 and nodewise ℓ_1-regularized linear regression (Neighborhood Lasso). As our theorem suggests and these figures show, the Neighborhood Greedy algorithm requires less samples to recover the exact structure of the graphical model.

$\tau \in (-1, 1)$, similar to our greedy methods.

5.3.3 Diamond Graph

Consider the diamond graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ on 4 nodes with the nodes labeled as in Fig 1(d). Given a correlation parameter $\tau \neq 0$, let Σ^* be the population covariance matrix with $\Sigma^*_{ii} = 1$ and $\Sigma^*_{ij} = \tau$ except $\Sigma^*_{23} = 0$ and $\Sigma^*_{14} = 2\tau^2$. It is easy to check that $\Sigma^{*\,-1}$ induces the desired graph. With this setup, the ℓ_1-regularized Gaussian MLE [10] requires $4|\tau|\,(|\tau| + 1) < 1$ or equivalently $\tau \in (-0.2017, 0.2017)$. Our greedy algorithm allows for $\tau \in (-0.7071, 0.7071)$ (since $C_{\min} = 1 - 2\tau^2$). Under the same setup, the linear neighborhood selection with ℓ_1-regularization [8] requires $2|\tau| < 1$ or equivalently that $\tau \in (-0.5, 0.5)$ to guarantee the success. Unlike the previous two examples, this is a strictly stronger condition than that imposed by our greedy methods.

6 Experimental Analysis

In this section we will outline our experimental results in testing the effectiveness of both Algorithms 1 and 2 in a simulated environment.

6.1 Optimization Method

Our greedy algorithm consists of a single variable optimization step where we try to pick the best coordinate. This step can be run in parallel for all single variables to achieve maximum speedup. For greedy neighborhood selection, the single variable optimization is a relatively simple operation, however for the global model selection algorithm (log-det optimization), we would like to provide a fast single variable optimization method to avoid a continual log-det calculation. Following the result in [12], we have

$$\det\left(\widehat{\Theta}^{(k-1)} + \alpha(e_{ij} + e_{ji})\right) = \det\left(\widehat{\Theta}^{(k-1)}\right) \left((1 + \alpha(\widehat{\Theta}^{(k-1)})^{-1}_{i,j})^2 - \alpha^2(\widehat{\Theta}^{(k-1)})^{-1}_{ii}(\widehat{\Theta}^{(k-1)})^{-1}_{jj}\right)$$

This entails that

$$\alpha^* = \arg\min_{\alpha} \langle\langle \widehat{\Theta}^{(k-1)} + \alpha(e_{ij} + e_{ji}),\ \widehat{\Sigma}^n \rangle\rangle - \log\det(\widehat{\Theta}^{(k-1)} + \alpha(e_{ij} + e_{ji}))$$
$$= \frac{\widehat{\Sigma}^n_{ij} - (\widehat{\Theta}^{(k-1)})^{-1}_{ij}}{(\widehat{\Theta}^{(k-1)})^{-1}_{ii}(\widehat{\Theta}^{(k-1)})^{-1}_{jj} - (\widehat{\Theta}^{(k-1)})^{-1}_{ij}(\widehat{\Theta}^{(k-1)})^{-1}_{ij}}$$

This closed-form solution simplifies the single variable optimization step in our algorithm and avoids continual calculation of $\log\det(\widehat{\Theta})$.

6.2 Experiments

To present a formal experimental analysis for both Algorithm 1 and Algorithm 2 we simulated zero-mean Gaussian inverse covariance estimation, or GMRF structure learning, for various graph types and scalings of (n, p, d). For the Global Greedy method we experimented using chain ($d = 2$) and grid ($d = 4$) graph types with sizes of $p \in \{36, 64, 100\}$. For the Neighborhood Greedy method we experimented using chain ($d = 2$) and star ($d = 0.1p$) graph types with sizes of $p \in \{36, 64, 100\}$. Figure 1 outlines the schematic structure for each graph type. For each algorithm, we measured performance by completely learning the true support set S^* pertaining to the non-zero inverse covariates (graph edges). If S^* was completely learned then we called this a *success* and otherwise we called it a *failure*. Using a batch size of 50 trials for each scaling of (n, p, d) we measured the probability of success as the average success rate. For both algorithms we used a stopping threshold $\epsilon_{\mathcal{S}} = \frac{cd\log p}{n}$ where d is the maximum degree of the graph, p is the number of nodes in the graph, n is the number of samples used, and c is a constant tuning parameter, as well as a backwards step threshold of $v = 0.5$. We compared Algorithm 1 to that of ℓ_1-regularized Gaussian MLE (Graphical Lasso) as discussed in [5] and [10] using the *glasso* implementation from Friedman et al. [5]. We compared Algorithm 2 to that of neighborhood based ℓ_1-regularized linear regression (Neighborhood Lasso) using the *glmnet* generalized Lasso implementation, also from Friedman et al. [6]. Both *glasso* and *glmnet* use a regularization parameter $\lambda = c\sqrt{\frac{logp}{n}}$ which was optimally set using k-fold cross validation.

Figure 2 plots the probability of successfully learning S^* vs the control parameter $\beta(n, p, d) = \frac{n}{70d\log p}$ for varying number of samples n for both Algorithm 1 and Graphical Lasso. Figure 3 plots the probability of successfully learning S^* vs the control parameter $\beta(n, p, d) = \frac{n}{70d\log p}$ for the chain graph type and $\beta(n, p, d) = \frac{n}{200\log(dp)}$ for the star graph type for both Algorithm 2 and neighborhood based ℓ_1-linear regression. Both figures illustrate our theoretical results that the Greedy Algorithms require less samples ($O(d\log p)$) than the state of the art Lasso methods ($O(d^2\log p)$) for complete structure learning.

References

[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[2] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

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

[4] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–374, 2000.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostat.*, 9(3):432–441, 2007.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.

[7] A. Jalali, C. Johnson, and P. Ravikumar. On learning discrete graphical models using greedy methods. In *NIPS*, 2011.

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

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

[10] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.*, 5:935–98, 2011.

[11] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. 2:494–515, 2008.

[12] K. Scheinberg and I. Rish. Learning sparse gaussian markov networks using a greedy coordinate ascent approach. *In: Balcazar, J., Bonchi, F., Gionis, A., Sebag, M. (eds.) Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science 6323*, pages 196–212.

[13] V. N. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235–409, 2008.

[14] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009.

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

[16] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *Neural Information Processing Systems (NIPS) 21*, 2008.

[17] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.