Structure Learning of Gaussian Markov Random Fields with False Discovery Rate Control

Sangkyun Lee [1],*[iD]**, Piotr Sobczyk** [2] **and Malgorzata Bogdan** [3]

[1]   Computer Science, Hanyang University ERICA, Ansan 15588, Korea
[2]   Department of Mathematics, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland;
     piotr.sobczyk@pwr.edu.pl
[3]   Institute of Mathematics, University of Wroclaw, 50-384 Wroclaw, Poland; Malgorzata.Bogdan@uwr.edu.pl
*   Correspondence: sangkyun@hanyang.ac.kr

check for
updates

Abstract: In this paper, we propose a new estimation procedure for discovering the structure of Gaussian Markov random fields (MRFs) with false discovery rate (FDR) control, making use of the sorted ℓ_1-norm (SL1) regularization. A Gaussian MRF is an acyclic graph representing a multivariate Gaussian distribution, where nodes are random variables and edges represent the conditional dependence between the connected nodes. Since it is possible to learn the edge structure of Gaussian MRFs directly from data, Gaussian MRFs provide an excellent way to understand complex data by revealing the dependence structure among many inputs features, such as genes, sensors, users, documents, etc. In learning the graphical structure of Gaussian MRFs, it is desired to discover the actual edges of the underlying but unknown probabilistic graphical model—it becomes more complicated when the number of random variables (features) p increases, compared to the number of data points n. In particular, when $p \gg n$, it is statistically unavoidable for any estimation procedure to include false edges. Therefore, there have been many trials to reduce the false detection of edges, in particular, using different types of regularization on the learning parameters. Our method makes use of the SL1 regularization, introduced recently for model selection in linear regression. We focus on the benefit of SL1 regularization that it can be used to control the FDR of detecting important random variables. Adapting SL1 for probabilistic graphical models, we show that SL1 can be used for the structure learning of Gaussian MRFs using our suggested procedure nsSLOPE (neighborhood selection Sorted L-One Penalized Estimation), controlling the FDR of detecting edges.

Keywords: Gaussian Markov random field; Inverse Covariance Matrix Estimation; FDR control

1. Introduction

Estimation of the graphical structure of Gaussian Markov random fields (MRFs) has been the topic of active research in machine learning, data analysis and statistics. The reason is that they provide efficient means for representing complex statistical relations of many variables in forms of a simple undirected graph, disclosing new insights about interactions of genes, users, news articles, operational parts of a human driver, to name a few.

One mainstream of the research is to estimate the structure by maximum likelihood estimation (MLE), penalizing the ℓ_1-norm of the learning parameters. In this framework, structure learning of a Gaussian MRF is equivalent to finding a sparse inverse covariance matrix of a multivariate Gaussian distribution. To formally describe the connection, let us we consider n samples $x_1, x_2, \ldots, x_n$ of p jointly Gaussian random variables following $\mathcal{N}(0, \Sigma_{p \times p})$, where the mean is zero without loss of generality and $\Sigma_{p \times p}$ is the covariance matrix. The estimation is essentially the MLE of the inverse

covariance matrix $\Theta := \Sigma^{-1}$ under an ℓ_1-norm penalty, which can be stated it as a convex optimization problem [1]:

$$\hat{\Theta} = \underset{\Theta \in \mathbb{R}^{p \times p}, \Theta' = \Theta}{\arg\min} \quad -\log\det\Theta + \operatorname{tr}(S\Theta) + \lambda\|\Theta\|_1 \tag{1}$$

Here, $S := \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ is the sample covariance matrix, $\|\Theta\|_1 := \sum_{i,j} |\Theta_{ij}|$ and $\lambda > 0$ is a tuning parameter that determines the element-wise sparsity of $\hat{\Theta}$.

The ℓ_1-regularized MLE approach (1) has been addressed quite extensively in literature. The convex optimization formulation has been first discussed in References [2,3]. A block-coordinate descent type algorithm was developed in Reference [4], while revealing the fact that the sub-problems of (1) can be solved in forms of the LASSO regression problems [5]. More efficient solvers have been developed to deal with high-dimensional cases [6–13]. In the theoretical side, we are interested in two aspects—one is the estimation quality of $\hat{\Theta}$ and another is the variable selection quality of $\hat{\Theta}$. These aspects are still in under investigation in theory and experiments [1,3,14–20], as new analyses become available for the closely related ℓ_1-penalized LASSO regression in vector spaces.

Among these, our method inherits the spirit of Reference [14,19] in particular, where the authors considered the problem (1) in terms of a collection of local regression problems defined for each random variables.

LASSO and SLOPE

Under a linear data model $b = A\beta + \epsilon$ with a data matrix $A \in \mathbb{R}^{n \times p}$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, the ℓ_1-penalized estimation of β is known as the LASSO regression [5], whose estimate is given by solving the following convex optimization problem:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \frac{1}{2}\|b - A\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda > 0$ is a tuning parameter which determines the sparsity of the estimate $\hat{\beta}$. Important statistical properties of $\hat{\beta}$ is that (i) the distance between the estimate $\hat{\beta}$ and the population parameter vector β and (ii) the detection of non-zero locations of β by $\hat{\beta}$. We can rephrase the former as the estimation error and the latter as the model selection error. These two types of error are dependent on how to choose the value of the tuning parameter λ. Regarding the model selection, it is known that the LASSO regression can control the family-wise error rate (FWER) at level $\alpha \in [0, 1]$ by choosing $\lambda = \sigma\Phi^{-1}(1 - \alpha/2p)$ [21]. The FWER is essentially the probability of including at least one entry as non-zero in $\hat{\beta}$ which is zero in β. In high-dimensional cases where $p \gg n$, controlling FWER is quite restrictive since it is unavoidable to do false positive detection of nonzero entries. As a result, FWER control can lead to weak detection power of nonzero entries.

The SLOPE is an alternative procedure for the estimation of β, using the sorted ℓ_1-norm penalization instead. The SLOPE solves a modified convex optimization problem,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \frac{1}{2}\|b - A\beta\|_2^2 + J_\lambda(\beta)$$

where $J_\lambda(\cdot)$ is the sorted ℓ_1-norm defined by

$$J_\lambda(\beta) := \sum_{i=1}^{p} \lambda_i |\beta|_{(i)}$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$ and $|\beta|_{(k)}$ is the kth largest component of β in magnitude. In Reference [21], it has been shown that, for linear regression, the SLOPE procedure can control the false discovery rate (FDR) at level $q \in [0, 1]$ of model selection by choosing $\lambda_i = \Phi^{-1}(1 - i \cdot q/2p)$. The FDR is the expected ratio of false discovery (i.e., the number of false nonzero entries) to total

discovery. Since controlling FDR is less restrictive for model selection compared to the FWER control, FDR control can lead to a significant increase in detection power, while it may slightly increase the total number of false discovery [21].

This paper is motivated by the SLOPE method [21] for its use of the SL1 regularization, where it brings many benefits not available with the popular ℓ_1-based regularization—the capability of false discovery rate (FDR) control [21,22], adaptivity to unknown signal sparsity [23] and clustering of coefficients [24,25]. Also, efficient optimization methods [13,21,26] and more theoretical analysis [23,27–29] are under active research.

In this paper, we propose a new procedure to find a sparse inverse covariance matrix estimate, we call nsSLOPE (neighborhood selection Sorted L-One Penalized Estimation). Our nsSLOPE procedure uses the sorted ℓ_1-norm for penalized model selection, whereas the existing gLASSO (1) method uses the ℓ_1-norm for the purpose. We investigate our method in two aspects in theory and in experiments, showing that (i) how the estimation error can be bounded, and (ii) how the model selection (more specifically, the neighborhood selection [14]) can be done with an FDR control in the edge structure of the Gaussian Markov random field. We also provide an efficient but straightforward estimation algorithm which fits for parallel computation.

2. nsSLOPE (Neighborhood Selection Sorted L-One Penalized Estimation)

Our method is based on the idea that the estimation of the inverse covariance matrix of a multivariate normal distribution $\mathcal{N}(0, \Sigma_{p \times p})$ can be decomposed into the multiple regression on conditional distributions [14,19].

For a formal description of our method, let us consider a p-dimensional random vector $x \sim \mathcal{N}(0, \Sigma_{p \times p})$, denoting its ith component as $x_i \in \mathbb{R}$ and the sub-vector without the ith component as $x_{-i} \in \mathbb{R}^{p-1}$. For the inverse covariance matrix $\Theta := \Sigma^{-1}$, we use Θ_i and Θ_{-i} to denote the ith column of the matrix and the rest of Θ without the ith column, respectively.

From the Bayes rule, we can decompose the full log-likelihood into the following parts:

$$\sum_{j=1}^{n} \log P_\Theta(x = x^j) = \sum_{j=1}^{n} \log P_{\Theta_i}(x_i = x_i^j | x_{-i} = x_{-i}^j)$$
$$+ \sum_{j=1}^{n} \log P_{\Theta_{-i}}(x_{-i} = x_{-i}^j).$$

This decomposition allows us a block-wise optimization of the full log-likelihood, which iteratively optimizes $P_{\Theta_i}(x_i | x_{-i})$ while the parameters in $P_{\Theta_{-i}}(x_{-i})$ are fixed at the current value.

2.1. Sub-Problems

In the block-wise optimization approach we mentioned above, we need to deal with the conditional distribution $P_{\Theta_i}(x_i | x_{-i})$ in each iteration. When $x \sim \mathcal{N}(0, \Sigma)$, the conditional distribution also follows the Gaussian distribution [30], in particular:

$$x_i | x_{-i} \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \begin{cases} \mu_i := x_{-i} \Sigma_{-i,-i}^{-1} \Sigma_{-i,i} \\ \sigma_i^2 := \Sigma_{ii} - \Sigma_{i,-i} \Sigma_{-i,-i}^{-1} \Sigma_{-i,i}. \end{cases}$$

Here $\Sigma_{-i,i} \in \mathbb{R}^{p-1}$ denotes the ith column of Σ without the ith row element, $\Sigma_{-i,-i} \in \mathbb{R}^{(p-1) \times (p-1)}$ denotes the sub-matrix of Σ without the ith column and the ith row and $\Sigma_{ii} = \Sigma_{i,i}$ is the ith diagonal component of Σ. Once we define β^i as follow,

$$\beta^i := \Sigma_{-i,-i}^{-1} \Sigma_{-i,i} \in \mathbb{R}^{p-1},$$

Symmetry **2019**, *11*, 1311
4 of 17

then the conditional distribution now looks like:

$$x_i|x_{-i} \sim N(\mu_i, \sigma_i^2), \quad \begin{cases} \mu_i = x_{-i}\beta^i \\ \sigma_i^2 = \Sigma_{ii} - (\beta^i)'\Sigma_{-i,-i}\beta^i. \end{cases}$$

This indicates that the minimization of the conditional log-likelihood can be understood as a local regression for the random variable x_i, under the data model:

$$x_i = x_{-i}\beta^i + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \sigma_i^2). \tag{2}$$

To obtain the solution of the local regression problem (2), we consider a convex optimization based on the sorted ℓ_1-norm [21,25]. In particular, for each local problem index $i \in \{1, \ldots, p\}$, we solve

$$\hat{\beta}^i \in \min_{\beta^i \in \mathbb{R}^{p-1}} \frac{1}{2}\|X_i - X_{-i}\beta^i\|_2^2 + \hat{\sigma}_i \cdot J_\lambda(\beta^i). \tag{3}$$

Here, the matrix $X \in \mathbb{R}^{n \times p}$ consists of n i.i.d. p-dimensional samples in rows, X_i is the ith column of X and $X_{-i} \in \mathbb{R}^{n \times (p-1)}$ is the sub-matrix of X without the ith column. Note that the sub-problem (3) requires an estimate of σ_i, namely $\hat{\sigma}_i$, which will be computed dependent on the information on the other sub-problem indices (namely, $\Sigma_{-i,-i}$). Because of this, the problems (3) for indices $i = 1, 2, \ldots, p$ are not independent. This contrasts our method to the neighborhood selection algorithms [14,19], based on the ℓ_1-regularization.

2.2. Connection to Inverse Covariance Matrix Estimation

With $\hat{\beta}^i$ obtained from (3) as an estimate of $\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}$ for $i = 1, 2, \ldots, p$, now the question is how to use the information for the estimation of $\Theta = \Sigma^{-1}$ without an explicit inversion of the matrix. For the purpose, we first consider the block-wise formulation of $\Sigma\Theta = I$:

$$\begin{bmatrix} \Sigma_{-i,-i} & \Sigma_{-i,i} \\ \Sigma_{i,-i} & \Sigma_{ii} \end{bmatrix} \begin{bmatrix} \Theta_{-i,-i} & \Theta_{-i,i} \\ \Theta_{i,-i} & \Theta_{ii} \end{bmatrix} = \begin{bmatrix} I_{(p-1)\times(p-1)} & 0 \\ 0 & 1 \end{bmatrix}$$

putting the ith row and the ith column to the last positions whenever necessary. There we can see that $\Sigma_{-i,-i}\Theta_{-i,i} + \Sigma_{-i,i}\Theta_{ii} = 0$. Also, from the block-wise inversion of $\Theta = \Sigma^{-1}$, we have (unnecessary block matrices are replaced with $*$):

$$\begin{bmatrix} \Sigma_{-i,-i} & \Sigma_{-i,i} \\ \Sigma_{i,-i} & \Sigma_{ii} \end{bmatrix}^{-1} = \begin{bmatrix} * & * \\ * & (\Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i})^{-1} \end{bmatrix}$$

From these and the definition of β^i, we can establish the following relations:

$$\begin{cases} \beta^i &= \Sigma_{-i,-i}^{-1}\Sigma_{-i,i} = -\frac{\Theta_{-i,i}}{\Theta_{ii}} \\ \Theta_{ii} &= (\Sigma_{ii} - (\beta^i)'\Sigma_{-i,-i}\beta^i)^{-1} = 1/\sigma_i^2. \end{cases} \tag{4}$$

Note that $\Theta_{ii} > 0$ for a positive definite matrix Θ and that (4) implies the sparsity pattern of $\Theta_{-i,i}$ and β^i must be the same. Also, the updates (4) for $i = 1, 2, \ldots, i$ are not independent, since the computation of $\sigma_i = 1/\Theta_{ii}$ depends on $\Sigma_{-i,-i}$. However, if we estimate σ_i based on the sample covariance matrix instead of Σ, (i) our updates (4) no longer needs to explicitly compute or store the $\Sigma_{p \times p}$ matrix, unlike the gLASSO and (ii) our sub-problems become mutually independent and thus solvable in parallel.

3. Algorithm

Our algorithm, called nsSLOPE, is summarized in Algorithm 1, which is essentially a block-coordinate descent algorithm [12,31]. Our algorithm may look similar to that of Yuan [19] but there are several important differences—(i) each sub-problem in our algorithm solves a SLOPE formulation (SL1 regularization), while Yuan's sub-problem is either LASSO or Dantzig selector (ℓ_1 regularization); (ii) our sub-problem makes use of the estimate $\hat{\Theta}_{ii}$ in addition to $\hat{\Theta}_{-i,i}$.

Algorithm 1: The nsSLOPE Algorithm

Input: $X \in \mathbb{R}^{n \times p}$ with zero-centered columns
Input: $S = \frac{1}{n} X'X$
Input: The target level of FDR q
Set $\lambda = (\lambda_1, \ldots, \lambda_{p-1})$ according to Section 3.2;
while $\hat{\Theta}$ *not converged* **do**
    for $i = 1, 2, \ldots, p$ **do**
        Compute $\hat{\beta}^i$ as the solution to the sub-problem (3):

$$\hat{\beta}^i = \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min} \ \frac{1}{2} \|X_i - X_{-i}\beta\|^2 + (\hat{\Theta}_{ii})^{-1/2} J_\lambda(\beta)$$

        Update $\hat{\Theta}_{ii}$ and $\hat{\Theta}_{-i,i}$ ((4) and Section 3.3):

$$\begin{cases} \hat{\Theta}_{ii} &= (S_{ii} - 2(\hat{\beta}^i)'S_{-i,i} + (\hat{\beta}^i)'S_{-i,-i}\hat{\beta}^i)^{-1} \\ \hat{\Theta}_{-i,i} &= -\hat{\Theta}_{ii}\hat{\beta}^i \end{cases}$$

    Make $\hat{\Theta}$ a symmetric matrix:

$$\hat{\Theta} \leftarrow \underset{\Theta = \Theta'}{\arg\min} \ \|\Theta - \hat{\Theta}\|_F^2$$

3.1. Sub-Problem Solver

To solve our sub-problems (3), we use the SLOPE R package, which implements the proximal gradient descent algorithm of Reference [32] with acceleration based on Nesterov's original idea [33]. The algorithm requires to compute the proximal operator involving $J_\lambda(\cdot)$, namely

$$\text{prox}_{J_\lambda}(z) := \underset{x \in \mathbb{R}^{p-1}}{\arg\min} \ \frac{1}{2}\|x - z\|^2 + J_\lambda(x).$$

This can be computed in $\mathcal{O}(p \log p)$ time using an algorithm from Reference [21]. The optimality of sub-problem is declared by the primal-dual gap and we use a tight threshold value, 10^{-7}.

3.2. Choice of $\lambda = (\lambda_1, \ldots, \lambda_{p-1})$

For the sequence of λ values in sub-problems, we use so-called the Benjamini-Hochberg (BH) sequence [21]:

$$\lambda_i^{BH} = \Phi^{-1}(1 - iq/2(p-1)), \ i = 1, \ldots, p-1. \tag{5}$$

Here $q \in [0, 1]$ is the target level of FDR control we discuss later and $\Phi^{-1}(\alpha)$ is the αth quantile of the standard normal distribution. In fact, when the design matrix X_{-i} in the SLOPE sub-problem (3)

is not orthogonal, it is beneficial to use an adjusted version of this sequence. This sequence is generated by:

$$\lambda_i = \lambda_i^{BH} \sqrt{1 + w(i-1) \sum_{j<i} \lambda_j^2}, \quad w(k) = \frac{1}{n-k-1},$$

and if the sequence $\lambda_1, \ldots, \lambda_k$ is non-increasing but λ_i begins to increase after the index k^*, we set all the remaining values equal to λ_{k^*}, so that the resulting sequence will be non-increasing. For more discussion about the adjustment, we refer to Section 3.2.2 of Reference [21].

3.3. Estimation of Θ_{ii}

To solve the ith sub-problem in our algorithm, we need to estimate the value of Θ_{ii}. This can be done using (4), that is, $\Theta_{ii} = (\Sigma_{ii} - (\beta^i)'\Sigma_{-i,-i}\beta^i)^{-1}$. However, this implies that (i) we need to keep an estimate of $\Sigma \in \mathbb{R}^{p \times p}$ additionally, (ii) the computation of the ith sub-problem will be dependent on all the other indices, as it needs to access $\Sigma_{-i,-i}$, requiring the algorithm to run sequentially.

To avoid these overheads, we compute the estimate using the sample covariance matrix $S = \frac{1}{n}X'X$ instead (we assume that the columns of $X \in \mathbb{R}^{n \times p}$ are centered):

$$\tilde{\Theta}_{ii} = (S_{ii} - 2\hat{\beta}'S_{-i,i} + \hat{\beta}'S_{-i,-i}\hat{\beta})^{-1}.$$

This allows us to compute the inner loop of Algorithm 1 in parallel.

3.4. Stopping Criterion of the External Loop

To terminate the outer loop in Algorithm 1, we check if the diagonal entries of $\hat{\Theta}$ have converged, that is, the algorithm is stopped when the ℓ_∞-norm different between two consecutive iterates is below a threshold value, 10^{-3}. The value is slightly loose but we have found no practical difference by making it tighter. Note that it is suffice to check the optimality of the diagonal entries of $\hat{\Theta}$ since the optimality of $\hat{\beta}^i$'s is enforced by the sub-problem solver and $\hat{\Theta}_{-i,i} = -\hat{\Theta}_{ii}\hat{\beta}^i$.

3.5. Uniqueness of Sub-Problem Solutions

When $p > n$, our sub-problems may have multiple solutions, which may prevent the global convergence of our algorithm. We may adopt the technique in Reference [34] to inject a strongly convex proximity term into each sub-problem objective, so that each sub-problem will have a unique solution. In our experience, however, we encountered no convergence issues using stopping threshold values in the range of $10^{-3} \sim 10^{-7}$ for the outer loop.

4. Analysis

In this section, we provide two theoretical results of our nsSLOPE procedure—(i) an estimation error bound regarding the distance between our estimate $\tilde{\Theta}$ and the true model parameter Θ and (ii) group-wise FDR control on discovering the true edges in the Gaussian MRF corresponding to Θ.

We first discuss about the estimation error bound, for which we divide our analysis into two parts regarding (i) off-diagonal entries and (ii) diagonal entries of Θ.

4.1. Estimation Error Analysis

4.1.1. Off-Diagonal Entries

From (4), we see that $\hat{\beta}^i = -\hat{\Theta}_{-i,i}/\hat{\Theta}_{ii}$, in other words, when $\hat{\Theta}_{ii}$ is fixed, the off-diagonal entries $\hat{\Theta}_{-i,i}$ is determined solely by $\hat{\beta}^i$, and therefore we can focus on the estimation error of $\hat{\beta}^i$.

To discuss the estimation error of $\hat{\beta}^i$, it is convenient to consider a constrained reformulation of the sub-problem (3):

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^d} J_\lambda(\beta) \quad \text{s.t.} \quad \frac{1}{n}\|b - A\beta\|_1 \le \epsilon. \tag{6}$$

Hereafter, for the sake of simplicity, we use notations $b := X_i \in \mathbb{R}^n$ and $A := X_{-i} \in R^{n \times d}$ for $d := p - 1$, also dropping sub-problem indices in $\hat{\beta}$ and $\epsilon > 0$. In this view, the data model being considered in each sub-problem is as follows,

$$b = A\beta^* + \nu, \quad \nu \sim \mathcal{N}(0, \sigma_i^2 I_n).$$

For the analysis, we make the following assumptions:

1. The true signal $\beta^* \in \mathbb{R}^d$ satisfies $\|\beta^*\|_1 \le \sqrt{s}$ for some $s > 0$ (this condition is satisfied for example, if $\|\beta^*\|_2 \le 1$ and β^* is s-sparse, that is, it has at most s nonzero elements).
2. The noise satisfies the condition $\frac{1}{n}\|\nu\|_1 \le \epsilon$. This will allow us to say that the true signal β^* is feasible with respect to the constraint in (6).

We provide the following result, which shows that $\hat{\beta}$ approaches the true β^* in high probability:

Theorem 1. *Suppose that $\hat{\beta}$ is an estimate of β^* obtained by solving the sub-problem (6). Consider the factorization $A := X_{-i} = BC$ where $B \in \mathbb{R}^{n \times k}$ ($n \le k$) is a Gaussian random matrix whose entries are sampled i.i.d. from $\mathcal{N}(0,1)$ and $C \in \mathbb{R}^{k \times d}$ is a deterministic matrix such that $C'C = \Sigma_{-i,-i}$. Such decomposition is possible since the rows of A are independent samples from $\mathcal{N}(0, \Sigma_{-i,-i})$. Then we have,*

$$\|\hat{\beta} - \beta^*\|_C^2 \le \sqrt{2\pi} \left(8\sqrt{2}\|C\|_1 \frac{\lambda_1}{\bar{\lambda}} \sqrt{\frac{s \log k}{n}} + \epsilon + t \right)$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2s}\frac{\bar{\lambda}^2}{\lambda_1^2}\right)$, where $\bar{\lambda} = \frac{1}{d}\sum_{i=1}^d \lambda_i$ and $\|C\|_1 = \max_{j=1,\dots,d}\sum_{i=1}^k |C_{ij}|$.

We need to discuss a few results before proving Theorem 1.

Theorem 2. *Let T be a bounded subset of $\mathbb{R}^d$. For an $\epsilon > 0$, consider the set*

$$T_\epsilon := \left\{ u \in T : \frac{1}{n}\|Au\|_1 \le \epsilon \right\}.$$

Then

$$\sup_{u \in T_\epsilon} (u'C'Cu)^{1/2} \le \sqrt{\frac{8\pi}{n}} \mathbb{E} \sup_{u \in S} |\langle C'g, u\rangle| + \sqrt{\frac{\pi}{2}}(\epsilon + t),$$

holds with probability at least $1 - 2\exp\left(-\frac{nt^2}{2d(T)^2}\right)$, where $g \in \mathbb{R}^k$ is a standard Gaussian random vector and $d(T) := \max_{u \in T} \|u\|_2$.

Proof. The result follows from an extended general M^* inequality in expectation [35]. □

The next result shows that the first term of the upper bound in Theorem 2 can be bounded without using expectation.

Lemma 1. *The quantity $\mathbb{E}\sup_{u \in \mathcal{K} - \mathcal{K}} |\langle C'g, u\rangle|$ is called the* width *of $\mathcal{K}$ and is bounded as follows,*

$$\mathbb{E}\sup_{u \in \mathcal{K} - \mathcal{K}} |\langle C'g, u\rangle| \le 4\sqrt{2}\|C\|_1 \frac{\lambda_1}{\bar{\lambda}} \sqrt{s \log k}.$$

Proof. This result is a part of the proof for Theorem 3.1 in Reference [35]. □

Using Theorem 2 and Lemma 1, we can derive a high probability error bound on the estimation from noisy observations,

$$b = A\beta^* + \nu, \quad \frac{1}{n}\|\nu\|_1 \le \epsilon,$$

where the true signal β^* belongs to a bounded subset $\mathcal{K} \in \mathbb{R}^d$. The following corollaries are straightforward extensions of Theorems 3.3 and 3.4 of Reference [35], given our Theorem 2 (so we skip the proofs).

Corollary 1. *Choose $\hat{\beta}$ to be any vector satisfying that*

$$\hat{\beta} \in \mathcal{K}, \quad and \quad \frac{1}{n}\|b - A\hat{\beta}\|_1 \le \epsilon.$$

Then,

$$\sup_{\beta \in \mathcal{K}}[(\hat{\beta} - \beta)'C'C(\hat{\beta} - \beta)]^{1/2}$$

$$\le \sqrt{2\pi}\left(8\sqrt{2}\|C\|_1\frac{\lambda_1}{\bar{\lambda}}\sqrt{\frac{s\log k}{n}} + \epsilon + t\right)$$

with probability at least $1 - 2\exp\left(-\frac{2nt^2}{d(T)^2}\right)$.

Now we show the error bound for the estimates we obtain by solving the optimization problem (6). For the purpose, we make use of the Minkowski functional of the set $\mathcal{K}$,

$$\|\beta\|_{\mathcal{K}} := \inf\{r > 0 : r\beta \in \mathcal{K}\}$$

If $\mathcal{K} \subset \mathbb{R}^p$ is a compact and origin-symmetric convex set with non-empty interior, then $\|\cdot\|_{\mathcal{K}}$ defines a norm in $\mathbb{R}^p$. Note that $\beta \in \mathcal{K}$ if and only if $\|\beta\|_{\mathcal{K}} \le 1$.

Corollary 2.

$$\hat{\beta} \in \argmin_{\beta \in \mathbb{R}^p} \|\beta\|_{\mathcal{K}} \quad subject \ to \quad \frac{1}{n}\|b - A\beta\|_1 \le \epsilon.$$

Then

$$\sup_{\beta \in \mathcal{K}}[(\hat{\beta} - \beta)'C'C(\hat{\beta} - \beta)]^{1/2}$$

$$\le \sqrt{2\pi}\left(8\sqrt{2}\|C\|_1\frac{\lambda_1}{\bar{\lambda}}\sqrt{\frac{s\log k}{n}} + \epsilon + t\right)$$

with probability at least $1 - 2\exp\left(-\frac{2nt^2}{d(T)^2}\right)$.

Finally, we show that solving the constrained form of the sub-problems (6) also satisfies essentially the same error bound in Corollary 2.

Proof of Theorem 1. Since we assumed that $\|\beta\|_1 \le \sqrt{s}$, we construct the subset $\mathcal{K}$ so that all vectors β with $\|\beta\|_1 \le \sqrt{s}$ will be contained in $\mathcal{K}$. That is,

$$\mathcal{K} := \{\beta \in \mathbb{R}^p : J_\lambda(\beta) \le \lambda_1\sqrt{s}\}.$$

This is a sphere defined in the SL1-norm $J_\lambda(\cdot)$: in this case, the Minkowski functional $\|\cdot\|_{\mathcal{K}}$ is proportional to $J_\lambda(\cdot)$ and thus the same solution minimizes both $J_\lambda(\cdot)$ and $\|\cdot\|_{\mathcal{K}}$.

Recall that $d(T) := \max_{u \in T} \|u\|_2$ and we choose $T = \mathcal{K} - \mathcal{K}$. Since $\|\beta\|_2 \leq \|\beta\|_1 \leq \frac{1}{\lambda} J_\lambda(\beta)$, for $\beta \in \mathcal{K}$ we have $\|\beta\|_2 \leq \frac{\lambda_1}{\lambda}\sqrt{s}$. This implies that $d(T) = 2\frac{\lambda_1}{\lambda}\sqrt{s}$. $\square$

4.1.2. Diagonal Entries

We estimate Θ_{ii} based on the residual sum of squares (RSS) as suggested by [19],

$$\hat{\sigma}_i^2 = (\hat{\Theta}_{ii})^{-1} = \|X_i - X'_{-i}\hat{\beta}^i\|^2/n$$
$$= S_{ii} - 2(\hat{\beta}^i)'S_{-i,i} + (\hat{\beta}^i)'S_{-i,-i}\hat{\beta}^i.$$

Unlike in Reference [19], we directly analyze the estimation error of $\hat{\Theta}_{ii}$ based on a chi-square tail bound.

Theorem 3. *For all small enough $\alpha > 0$ so that $\alpha/\sigma_i^2 \in [0, 1/2)$, we have*

$$\mathbb{P}(|\hat{\Theta}_{ii}^{-1} - \Theta_{ii}^{-1} - \delta((\beta^i)^*, \hat{\beta}^i)| \geq \alpha) \leq \exp\left(-\frac{3}{16}n\alpha^2\right)$$

where for $\nu \sim \mathcal{N}(0, \sigma_i^2 I_n)$,

$$\delta(\beta^*, \hat{\beta}) := (\beta^* - \hat{\beta})'S_{-i,-i}(\beta^* - \hat{\beta}) + 2\nu'X_{-i}(\beta^* - \hat{\beta})/n$$

Proof. Using the same notation as the previous section, that is, $b = X_i$ and $A = X_{-i}$, consider the estimate in discussion,

$$\hat{\sigma}_i^2 = \|b - A\hat{\beta}\|^2/n = \|A(\beta^* - \hat{\beta}) + \nu\|^2/n$$

where the last equality is from $b = A\beta^* + \nu$. Therefore,

$$\hat{\sigma}_i^2 = (\beta^* - \hat{\beta})'S_{-i,-i}(\beta^* - \hat{\beta}) + 2\nu'A(\beta^* - \hat{\beta})/n + \nu'\nu/n$$

where $S_{-i,-i} := A'A/n = X'_{-i}X_{-i}/n$. The last term is the sum of squares of independent $\nu_i \sim \mathcal{N}(0, \sigma_i^2)$ and therefore it follows the chi-square distribution, that is, $(\nu'\nu)/\sigma_i^2 \sim \chi_n^2$. Applying the tail bound [36] for a chi-square random variable Z with d degrees of freedom:

$$\mathbb{P}(|d^{-1}Z - 1| \geq \alpha) \leq \exp\left(-\frac{3}{16}d\alpha^2\right), \quad \alpha \in [0, 1/2).$$

we get for all small enough $\alpha > 0$,

$$\mathbb{P}(|\nu'\nu/n - \sigma_i^2| \geq \alpha) \leq \exp\left(-\frac{3}{16}n\alpha^2\right).$$

$\square$

4.1.3. Discussion on Asymptotic Behaviors

Our two main results above, Theorems 1 and 3, indicate how well our estimate of the off-diagonal entries $\hat{\Theta}_{-i,i}$ and diagonal entries $\hat{\Theta}_{ii}$ would behave. Based on these results, we can discuss the estimation error of the full matrix $\hat{\Theta}$ compared to the true precision matrix Θ.

From Theorem 1, we can deduce that with $C'C = \Sigma_{-i,-i}$ and $z(n) = O(\sqrt{s \log k/n})$,

$$\text{ev}_{\min}(\Sigma_{-i,-i})\|\hat{\beta} - \beta^*\| \leq [(\hat{\beta} - \beta^*)\Sigma_{-i,-i}(\hat{\beta} - \beta)]^{1/2}$$
$$\leq \sqrt{2\pi}(z(n) + \epsilon + t)$$

where $\text{ev}_{\min}(\Sigma_{-i,-i})$ is the smallest eigenvalue of the symmetric positive definite $\Sigma_{-i,-i}$. That is, using the interlacing property of eigenvalues, we have

$$\|\hat{\beta} - \beta^*\| \leq \frac{\sqrt{2\pi}(z(n) + \epsilon + t)}{\text{ev}_{\min}(\Sigma_{-i,-i})} \leq \sqrt{2\pi}\|\Theta\|(z(n) + \epsilon + t)$$

where $\|\Theta\| = \text{ev}_{\max}(\Theta)$ is the spectral radius of Θ. Therefore when $n \to \infty$, the distance between $\hat{\beta}$ and β^* is bounded by $\sqrt{2\pi}\|\Theta\|(\epsilon + t)$. Here, we can consider $t \to 0$ in a way such as $nt^2 \to \infty$ as n increases, so that the bound in Theorem 1 will hold with the probability approaching one. That is, in rough asymptotics,

$$\|\hat{\beta} - \beta^*\| \leq \epsilon\sqrt{2\pi}\|\Theta\|. \tag{7}$$

Under the conditions above, Theorem 3 indicates that:

$$\delta(\beta^*, \hat{\beta}) \leq \|S\|\|\beta^* - \hat{\beta}\|^2 + 2\|v/n\|_1\|X_{-i}(\beta^* - \hat{\beta})\|_\infty$$
$$\leq 2\epsilon^2\|\Theta\|(\pi\|S\|\|\Theta\| + \sqrt{2\pi}\|X_{-i}\|_\infty)$$

using our assumption that $\|v/n\|_1 \leq \epsilon$. We can further introduce assumptions on $\|\Theta\|$ and $S\Theta$ as in Reference [19], so that we can quantify the upper bound but here we will simply say that $\delta(\beta^*, \hat{\beta}) \leq 2c\epsilon^2$, where c is a constant depending on the properties of the full matrices S and Θ. If this is the case, then from Theorem 3 for an $\alpha \to 0$ such that $n\alpha^2 \to \infty$, we see that

$$|\hat{\Theta}_{ii}^{-1} - \Theta_{ii}^{-1}| \leq \delta((\beta^i)^*, \hat{\beta}^i) \leq 2c\epsilon^2, \tag{8}$$

with the probability approaching one.

Therefore, if we can drive $\epsilon \to 0$ as $n \to \infty$, then (7) and (8) imply the convergence of the diagonal $\hat{\Theta}_{ii} \to \Theta_{ii}$ and the off-diagonal $\hat{\Theta}_{-i,i} \to \Theta_{-i,i}$ in probability. Since $\|X_i - X_{-i}\beta\|_1 \leq \epsilon$ (or $\|X_i - X_{-i}\beta^i\| \leq \epsilon_i$ more precisely), $\epsilon \to 0$ is likely to happen when $(p-1)/n \geq 1$, implying that $p \to \infty$ as well as $n \to \infty$.

4.2. Neighborhood FDR Control under Group Assumptions

Here we consider $\hat{\beta}$ obtained by solving the unconstrained form (3) with the same data model we discussed above for the sub-problems:

$$b = A\beta + v, \quad v \sim \mathcal{N}(0, \sigma^2 I_n) \tag{9}$$

with $b = X_i \in \mathbb{R}^n$ and $A = X_{-i} \in \mathbb{R}^{n \times d}$ with $d = p - 1$. Here we focus on a particular but an interesting case where the columns of A form orthogonal groups, that is, under the decomposition $A = BC$, $C'C = \Sigma_{-i,-i}$ forms a block-diagonal matrix. We also assume that the columns of A belonging to the same group are highly correlated, in the sense that for any columns a_i and a_j of A corresponding to the same group, their correlation is high enough to satisfy that

$$\|a_i - a_j\| \leq \min_{i=1,\ldots,d-1}\{\lambda_i - \lambda_{i+1}\}/\|b\|.$$

This implies that $\hat{\beta}_i = \hat{\beta}_j$ by Theorem 2.1 of Reference [25], which simplifies our analysis. Note that if a_i and a_j belong to different blocks, then our assumption above implies $a_i' a_j = 0$. Finally, we further assume that $\|a_i\| \leq 1$ for $i = 1, \dots, d$.

Consider a collection of non-overlapping index subsets $g \subset \{1, \dots, d\}$ as the set of groups G. Under the block-diagonal covariance matrix assumption above, we see that

$$A\beta = \sum_{g \in G} \sum_{i \in g} a_i \beta_i = \sum_{g \in G} \left(\sum_{i \in g} a_i \right) \beta_g = \sum_{g \in G} \frac{\tilde{a}_g}{\|\tilde{a}_g\|} \beta_g'$$

where β_g denotes the representative of the same coefficients β_i for all $i \in g$, $\tilde{a}_g := \sum_{i \in g} a_i$ and $\beta_g' := \|\tilde{a}_g\| \beta_g$. This tells us that we can replace $A\beta$ by $\tilde{A}\beta'$, if we define $\tilde{A} \in \mathbb{R}^{n \times |G|}$ containing $\frac{\tilde{a}_g}{\|\tilde{a}_g\|}$ in its columns (so that $\tilde{A}'\tilde{A} = I_{|G|}$) and consider the vector of group-representative coefficients $\beta' \in \mathbb{R}^{|G|}$.

The regularizer can be rewritten similarly,

$$J_\lambda(\beta) = \sum_{i=1}^{|G|} \lambda_i' |\beta'|_{(i)} = J_{\lambda'}(\beta')$$

where $\lambda_i' := (\lambda_{\sum_{j=1}^{i-1} |g_{(j)}|+1} + \cdots + \lambda_{\sum_{j=1}^{i} |g_{(i)}|}) / \|\tilde{a}_{g_{(i)}}\|$, denoting by $g_{(i)}$ the group which has the ith largest coefficient in β' in magnitude and by $|g_{(i)}|$ the size of the group.

Using the fact that $\tilde{A}'\tilde{A} = I_{|G|}$, we can recast the regression model (9) as $b' = \tilde{A}'b = \beta + \tilde{A}'\nu \sim \mathcal{N}(\beta, \sigma^2 I_{|G|})$, and consider a much simpler form of the problem (3),

$$\hat{\beta}' = \arg\min_{\beta' \in \mathbb{R}^{|G|}} \frac{1}{2} \|b' - \beta'\|^2 + \sigma J_{\lambda'}(\beta'), \tag{10}$$

where we can easily check that $\lambda' = (\lambda_1', \dots, \lambda_{|G|}')$ satisfies $\lambda_1' \geq \cdots \geq \lambda_{|G|}'$. This is exactly the form of the SLOPE problem with an orthogonal design matrix in Reference [21], except that the new λ' sequence is not exactly the Benjamini-Hochberg sequence (5).

We can consider the problem (3) (and respectively (10)) in the context of multiple hypothesis testing of d (resp. $|G|$) null hypothesis $H_i : \beta_i = 0$ (resp. $H_{g_i} : \beta_i' = 0$) for $i = 1, \dots, d$ (resp. $i = 1, \dots, |G|$), where we reject H_i (resp. H_{g_i}) if and only if $\hat{\beta}_i \neq 0$ (resp. $\hat{\beta}_i' \neq 0$). In this setting, the Lemmas B.1 and B.2 in Reference [21] still holds for our problem (10), since they are independent of choosing the particular λ sequence.

In the following, V is the number of individual false rejections, R is the number of total individual rejections and R_g is the number of total group rejections. The following lemmas are slightly modified versions of Lemmas B.1 and B.2 from Reference [21], respectively, to fit our group-wise setup.

Lemma 2. *Let H_g be a null hypothesis and let $r \geq 1$. Then*

$$\{b' : H_g \text{ is rejected } \wedge R_g = r\} = \{b' : |b_g'| > \sigma \lambda_r' \wedge R_g = r\}.$$

Lemma 3. *Consider applying the procedure (10) for new data $\tilde{b}' = (b_1', \dots, b_{g-1}', b_{g+1}', \dots, b_{|G|}')$ with $\tilde{\lambda} = (\lambda_2', \dots, \lambda_d')$ and let $\tilde{R}_g$ be the number of group rejections made. Then with $r \geq 1$,*

$$\{b' : |b_g'| > \sigma \lambda_r' \wedge R_g = r\} \subset \{b' : |b_g'| > \sigma \lambda_r' \wedge \tilde{R}_g = r - 1\}.$$

Using these, we can show our FDR control result.

Theorem 4. *Consider the procedure (10) we consider in a sub-problem of nsSLOPE as a multiple testing of group-wise hypotheses, where we reject the null group hypothesis $H_g : \beta_g' = 0$ when $\beta_g' \neq 0$, rejecting all*

individual hypotheses in the group, that is, all H_i, $i \in g$, are rejected. Using $\lambda_1, \ldots, \lambda_{p-1}$ defined as in (5), the procedure controls FDR at the level $q \in [0, 1]$:

$$FDR = \mathbb{E}\left[\frac{V}{R \vee 1}\right] \leq \frac{G_0 q}{d} \leq q$$

where

$$
\begin{cases}
G_0 & := |\{g : (\beta'_g)^* = 0\}| \quad (\text{# true null group hypotheses}) \\
V & := |\{i : \beta_i^* = 0, \hat{\beta}_i^* \neq 0\}| \quad (\text{# false ind. rejections}) \\
R & := |\{i : \hat{\beta}_i^* \neq 0\}| \quad (\text{# all individual rejections})
\end{cases}
$$

Proof. Suppose that H_g is rejected. Then,

$$\mathbb{P}(H_g \text{ rejected } \wedge R_g = r) \stackrel{(i)}{\leq} \mathbb{P}(|b'_g| \geq \sigma \lambda'_r \wedge \tilde{R}_g = r - 1)$$

$$\stackrel{(ii)}{=} \mathbb{P}(|b'_g| \geq \sigma \lambda'_r)\, \mathbb{P}(\tilde{R}_g = r - 1)$$

$$\stackrel{(iii)}{\leq} \mathbb{P}\left(|b'_g| \geq \sigma \frac{|g_{(r)}| \lambda_{\sum_{j=1}^r |g_{(j)}|}}{\|\tilde{a}_{g_{(r)}}\|}\right) \mathbb{P}(\tilde{R}_g = r - 1)$$

$$\stackrel{(iv)}{\leq} \mathbb{P}(|b'_g| \geq \sigma \lambda_{\sum_{j=1}^r |g_{(j)}|})\, \mathbb{P}(\tilde{R}_g = r - 1)$$

The derivations above are—(i) by using Lemmas 2 and 3; (ii) from the independence between b'_g and $\tilde{b}'$; (iii) by taking the smallest term in the summation of λ'_r multiplied by the number of terms; (iv) due to the assumption that $\|a_i\| \leq 1$ for all i, so that $\|\tilde{a}_{g_{(r)}}\| \leq |g_{(r)}|$ by triangle inequality.

Now, consider the group-wise hypotheses testing in (10) is configured in a way that the first G_0 hypotheses are null in truth, that is, $H_{g_i} : \beta'_i = 0$ for $i \leq G_0$. Then we have:

$$FDR = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \sum_{r=1}^{|G|} \mathbb{E}\left[\frac{V}{\sum_{j=1}^r |g_{(j)}|} \mathbb{1}_{\{R_g = r\}}\right]$$

$$= \sum_{r=1}^{|G|} \frac{1}{\sum_{j=1}^r |g_{(j)}|} \sum_{i=1}^{G_0} \mathbb{E}[\mathbb{1}_{\{H_{g_i} \text{ rejected}\}} \mathbb{1}_{\{R_{g_i} = r\}}]$$

$$= \sum_{r=1}^{|G|} \frac{1}{\sum_{j=1}^r |g_{(j)}|} \sum_{i=1}^{G_0} \mathbb{P}(H_{g_i} \text{ rejected} \wedge R_{g_i} = r)$$

$$\leq \sum_{r=1}^{|G|} \frac{1}{\sum_{j=1}^r |g_{(j)}|} \sum_{i=1}^{G_0} \frac{\sum_{j=1}^r |g_{(j)}| q}{d} \mathbb{P}(\tilde{R}_{g_i} = r - 1)$$

$$= \sum_{r \geq 1} q \frac{G_0}{d} \mathbb{P}(\tilde{R}_{g_i} = r - 1) = q \frac{G_0}{d} \leq q.$$

□

Since the above theorem applies for each sub-problem, which can be considered as for the *i*th random variable to find its neighbors to be connected in the Gaussian Markov random field defined by Θ, we call this result as neighborhood FDR control.

5. Numerical Results

We show that the theoretical properties discussed above also works in simulated settings.

5.1. Quality of Estimation

For all numerical examples here, we generate $n = 100, 200, 300, 400$ random i.i.d. samples from $\mathcal{N}(0, \Sigma_{p \times p})$, where $p = 500$ is fixed. We plant a simple block diagonal structure into the true matrix Σ, which is also preserved in the precision matrix $\Theta = \Sigma^{-1}$. All the blocks had the same size of 4, so that we have total 125 blocks at the diagonal of Θ. We set $\Theta_{ii} = 1$ and set the entries Θ_{ij}, $i \neq j$, to a high enough value whenever (i, j) belongs to those blocks, in order to represent groups of highly correlated variables. All experiments were repeated for 25 times.

The λ sequence for nsSLOPE has been chosen according to Section 3.2 with respect to a target FDR level $q = 0.05$. For the gLASSO, we have used the λ value discussed in Reference [3], to control the family-wise error rate (FWER, the chance of any false rejection of null hypotheses) to $\alpha = 0.05$.

5.1.1. Mean Square Estimation Error

Recall our discussion of error bounds in Theorem 1 (off-diagonal scaled by $\hat{\Theta}_{ii}$) and Theorem 3 (diagonal), followed by Section 4.1.3 where we roughly sketched the asymptotic behaviors of $\hat{\Theta}_{ii}$ and $\hat{\Theta}_{-i,i}$.

The top panel of Figure 1 shows mean square error (MSE) between estimated quantities and the true models that we have created, where estimates are obtained by our method nsSLOPE, without or with symmetrization at the end of Algorithm 1, as well as by the gLASSO [4] which solves the ℓ_1-based MLE problem (1) with a block coordinate descent strategy. The off-diagonal estimation was consistently good overall settings, while the estimation error of diagonal entries were kept improving as n is being increased, which was we predicted in Section 4.1.3. We believe that our estimation of the diagonal has room for improvement, for example, using more accurate reverse-scaling to compensate the normalization within the SLOPE procedure.

5.1.2. FDR Control

A more exciting part of our results is the FDR control discussed in Theorem 4 and here we check how well the sparsity structure of the precision matrix is recovered. For comparison, we measure the power (the fraction of the true nonzero entries discovered by the algorithm) and the FDR (for the whole precision matrix).

The bottom panel of Figure 1 shows the statistics. In all cases, the empirical FDR was controlled around the desired level 0.05 by all methods, although our method kept the level quite strictly, having significantly larger power than the gLASSO. This is understandable since FWER control of the gLASSO if often too respective, thereby limiting the power to detect true positive entries. It is also consistent with the results reported for SL1-based penalized regression [21,23], which indeed is one of the key benefits of SL1-based methods.

5.2. Structure Discovery

To further investigate the structure discovery by nsSLOPE, we have experimented with two different types of covariance matrices—one with a block-diagonal structure and another with a hub structure. The covariance matrix of the block diagonal case has been constructed using the data.simulation function from the `varclust` R package, with $n = 100$, $SNR = 1$, $K = 4$, $numb.vars = 10$, $max.dim = 3$ (this gives us a data matrix with 100 examples and 40 variables). In the hub structure case, we have created a 20-dimensional covariance matrix with ones on the diagonal and 0.2 in the first column and the last row of the matrix. Then we have used the mvrnorm function from the `MASS` R package to sample 500 data points from a multivariate Gaussian distribution with zero mean and the constructed covariance matrix.
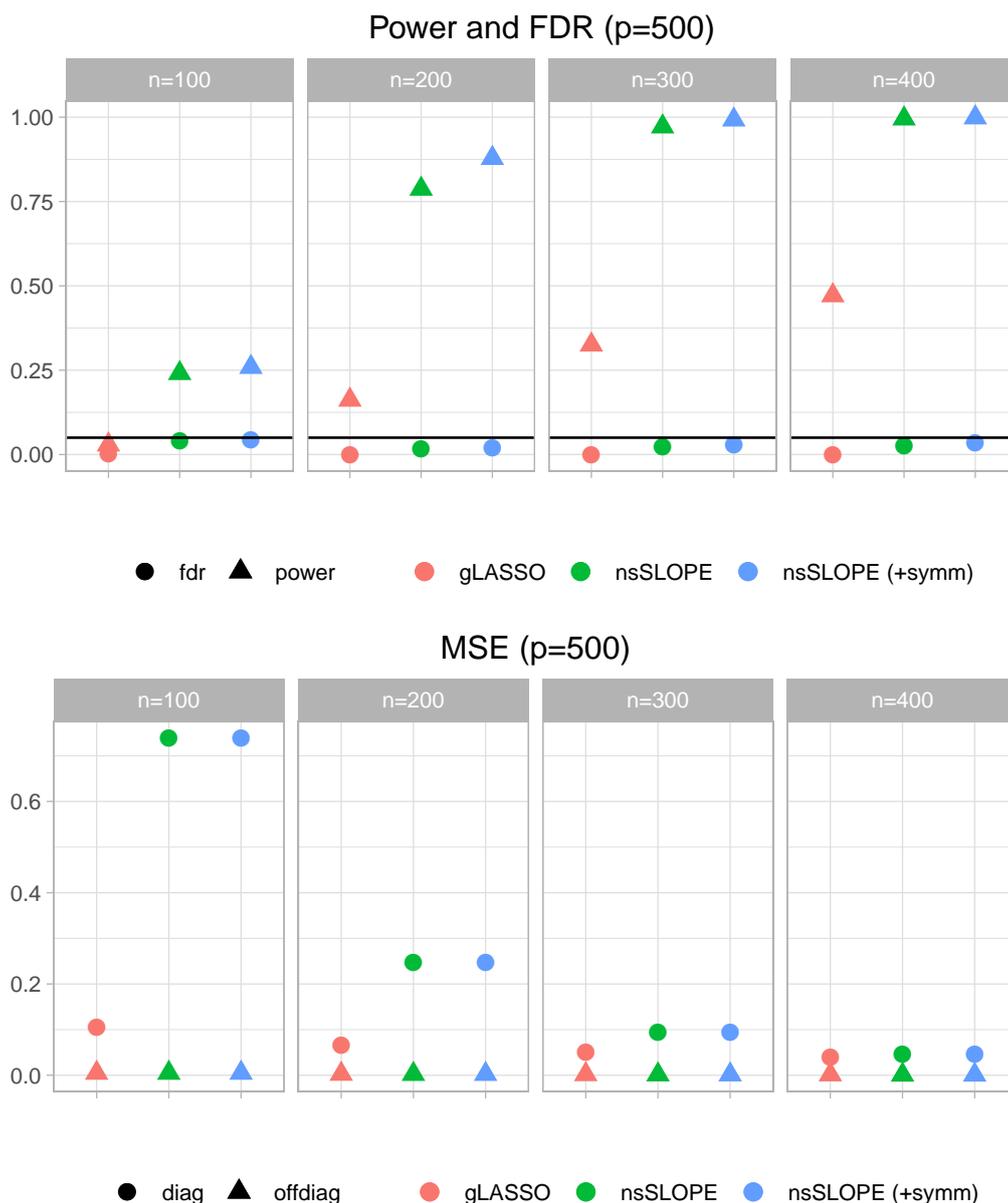
Figure 1. Quality of estimation. Top: empirical false discovery rate (FDR) levels (averaged over 25 repetitions) and the nominal level of $q = 0.05$ (solid black horizontal line). Bottom: mean square error of diagonal and off-diagonal entries of the precision matrix. $p = 500$ was fixed for both panels and $n = 100, 200, 300$ and 400 were tried. ("nsSLOPE": nsSLOPE without symmetrization, "+symm": with symmetrization and gLASSO.)

The true covariance matrix and the two estimates from gLASSO and nsSLOPE are shown in Figure 2. In nsSLOPE, the FDR control has been used based on the Benjamini-Hochberg sequence (5) with a target FDR value $q = 0.05$ and in gLASSO the FWER control [3] has been used with a target value $\alpha = 0.05$. Our method nsSLOPE appears to be more sensitive for finding the true structure, although it may contain a slightly more false discovery.
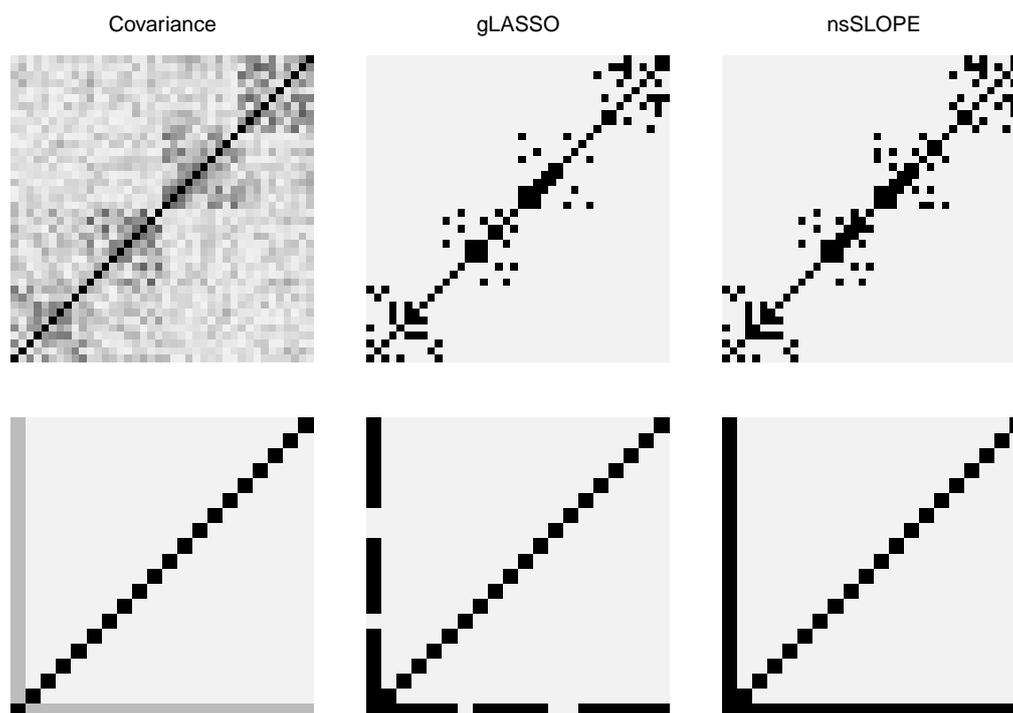
Figure 2. Examples of structure discovery. Top: a covariance matrix with block diagonal structure. Bottom: a hub structure. True covariance matrix is shown on the left and gLASSO and nsSLOPE estimates (only the nonzero patterns) of the precision matrix are shown in the middle and in the right panels, respectively.

6. Conclusions

We introduced a new procedure based on the recently proposed sorted ℓ_1 regularization, to find solutions of sparse precision matrix estimation with more attractive statistical properties than the existing ℓ_1-based frameworks. We believe there are many aspects of SL1 in graphical models to be investigated, especially when the inverse covariance has a more complex structure. Still, we hope our results will provide a basis for research and practical applications.

Our selection of the λ values in this paper requires independence assumptions on features or blocks of features. Although some extensions are possible [21], it would be desirable to consider a general framework, for example, based on Bayesian inference considering the posterior distribution derived from the loss and the regularizer [37,38], which enables us to evaluate the uncertainty of edge discovery and to find λ values from data.

References

1. Yuan, M.; Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* **2007**, *94*, 19–35. [CrossRef]
2. D'Aspremont, A.; Banerjee, O.; El Ghaoui, L. First-Order Methods for Sparse Covariance Selection. *SIAM J. Matrix Anal. Appl.* **2008**, *30*, 56–66. [CrossRef]
3. Banerjee, O.; Ghaoui, L.E.; d'Aspremont, A. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *J. Mach. Learn. Res.* **2008**, *9*, 485–516.

4. Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432–441. [CrossRef] [PubMed]

5. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. (Ser. B)* **1996**, *58*, 267–288. [CrossRef]

6. Oztoprak, F.; Nocedal, J.; Rennie, S.; Olsen, P.A. Newton-Like Methods for Sparse Inverse Covariance Estimation. In *Advances in Neural Information Processing Systems 25*; MIT Press: Cambridge, MA, USA, 2012; pp. 764–772.

7. Rolfs, B.; Rajaratnam, B.; Guillot, D.; Wong, I.; Maleki, A. Iterative Thresholding Algorithm for Sparse Inverse Covariance Estimation. In *Advances in Neural Information Processing Systems 25*; MIT Press: Cambridge, MA, USA, 2012; pp. 1574–1582.

8. Hsieh, C.J.; Dhillon, I.S.; Ravikumar, P.K.; Sustik, M.A. Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation. In *Advances in Neural Information Processing Systems 24*; MIT Press: Cambridge, MA, USA, 2011; pp. 2330–2338.

9. Hsieh, C.J.; Banerjee, A.; Dhillon, I.S.; Ravikumar, P.K. A Divide-and-Conquer Method for Sparse Inverse Covariance Estimation. In *Advances in Neural Information Processing Systems 25*; MIT Press: Cambridge, MA, USA, 2012; pp. 2330–2338.

10. Hsieh, C.J.; Sustik, M.A.; Dhillon, I.; Ravikumar, P.; Poldrack, R. BIG & QUIC: Sparse Inverse Covariance Estimation for a Million Variables. In *Advances in Neural Information Processing Systems 26*; MIT Press: Cambridge, MA, USA, 2013; pp. 3165–3173.

11. Mazumder, R.; Hastie, T. Exact Covariance Thresholding into Connected Components for Large-scale Graphical Lasso. *J. Mach. Learn. Res.* **2012**, *13*, 781–794.

12. Treister, E.; Turek, J.S. A Block-Coordinate Descent Approach for Large-scale Sparse Inverse Covariance Estimation. In *Advances in Neural Information Processing Systems 27*; MIT Press: Cambridge, MA, USA, 2014; pp. 927–935.

13. Zhang, R.; Fattahi, S.; Sojoudi, S. *Large-Scale Sparse Inverse Covariance Estimation via Thresholding and Max-Det Matrix Completion*; International Conference on Machine Learning, PMLR: Stockholm, Sweden, 2018.

14. Meinshausen, N.; Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* **2006**, *34*, 1436–1462. [CrossRef]

15. Meinshausen, N.; Bühlmann, P. Stability selection. *J. R. Stat. Soc. (Ser. B)* **2010**, *72*, 417–473. [CrossRef]

16. Rothman, A.J.; Bickel, P.J.; Levina, E.; Zhu, J. Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2008**, *2*, 494–515. [CrossRef]

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

18. Raskutti, G.; Yu, B.; Wainwright, M.J.; Ravikumar, P.K. Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems 21*; MIT Press: Cambridge, MA, USA, 2009; pp. 1329–1336.

19. Yuan, M. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *J. Mach. Learn. Res.* **2010**, *11*, 2261–2286.

20. Fattahi, S.; Zhang, R.Y.; Sojoudi, S. Sparse Inverse Covariance Estimation for Chordal Structures. In Proceedings of the 2018 European Control Conference (ECC), Limassol, Cyprus, 12–15 June 2018; pp. 837–844.

21. Bogdan, M.; van den Berg, E.; Sabatti, C.; Su, W.; Candes, E.J. SLOPE—Adaptive Variable Selection via Convex Optimization. *Ann. Appl. Stat.* **2015**, *9*, 1103–1140. [CrossRef] [PubMed]

22. Brzyski, D.; Su, W.; Bogdan, M. Group SLOPE—Adaptive selection of groups of predictors. *arXiv* **2015**, arXiv:1511.09078.

23. Su, W.; Candès, E. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Stat.* **2016**, *44*, 1038–1068. [CrossRef]

24. Bondell, H.D.; Reich, B.J. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics* **2008**, *64*, 115–123. [CrossRef]

25. Figueiredo, M.A.T.; Nowak, R.D. Ordered Weighted L1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, 9–11 May 2016; pp. 930–938.

26. Lee, S.; Brzyski, D.; Bogdan, M. Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l1-Norm. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain, 9–11 May 2016; Volume 51, pp. 780–789.

27. Chen, S.; Banerjee, A. Structured Matrix Recovery via the Generalized Dantzig Selector. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 3252–3260.

28. Bellec, P.C.; Lecué, G.; Tsybakov, A.B. Slope meets Lasso: Improved oracle bounds and optimality. *arXiv* **2017**, arXiv:1605.08651v3.

29. Derumigny, A. Improved bounds for Square-Root Lasso and Square-Root Slope. *Electron. J. Stat.* **2018**, *12*, 741–766. [CrossRef]

30. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*; Wiley-Interscience: London, UK, 2003.

31. Beck, A.; Tetruashvili, L. On the Convergence of Block Coordinate Descent Type Methods. *SIAM J. Optim.* **2013**, *23*, 2037–2060. [CrossRef]

32. Beck, A.; Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]

33. Nesterov, Y. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. *Soviet Math. Dokl.* **1983**, *27*, 372–376.

34. Razaviyayn, M.; Hong, M.; Luo, Z.Q. A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization. *SIAM J. Optim.* **2013**, *23*, 1126–1153. [CrossRef]

35. Figueiredo, M.; Nowak, R. Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. *arXiv* **2014**, arXiv:1409.4005.

36. Johnstone, I.M. Chi-square oracle inequalities. *Lect. Notes-Monogr. Ser.* **2001**, *36*, 399–418.

37. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]

38. Mallick, H.; Yi, N. A New Bayesian Lasso. *Stat. Its Interface* **2014**, *7*, 571–582. [CrossRef]