High-dimensional covariance estimation based on Gaussian graphical models

Shuheng Zhou SHUHENGZ@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1041, USA

Philipp Rütimann RUTIMANN@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
8092 Zürich, Switzerland

Min Xu MINX@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213-3815, USA

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH
Seminar for Statistics
ETH Zürich
8092 Zürich, Switzerland

Editor:

Abstract

Undirected graphs are often used to describe high dimensional distributions. Under sparsity conditions, the graph can be estimated using ℓ_1-penalization methods. We propose and study the following method. We combine a multiple regression approach with ideas of thresholding and refitting: first we infer a sparse undirected graphical model structure via thresholding of each among many ℓ_1-norm penalized regression functions; we then estimate the covariance matrix and its inverse using the maximum likelihood estimator. We show under suitable conditions that this approach yields consistent estimation in terms of graphical structure and fast convergence rates with respect to the Frobenius norm for the covariance matrix and its inverse. We also derive an explicit bound for the Kullback Leibler divergence.

Keywords: Graphical model selection, covariance estimation, Lasso, nodewise regression, thresholding

1. Introduction

There have been a lot of recent activities for estimation of high-dimensional covariance and inverse covariance matrices where the dimension p of the matrix may greatly exceed the sample size n. High-dimensional covariance estimation can be classified into two main categories, one which relies on a natural ordering among the variables Wu and Pourahmadi (2003); Bickel and Levina (2004); Huang et al. (2006); Furrer and Bengtsson (2007); Bickel and Levina (2008); Levina et al. (2008) and one where no natural ordering is given and estimators are permutation invariant with respect to indexing the variables Yuan and Lin (2007); Friedman et al. (2007); d'Aspremont et al. (2008); Banerjee et al. (2008); Rothman et al. (2008). We focus here on the latter class with permutation invariant estimation and we aim for an estimator which is accurate for both the covariance matrix Σ and its inverse, the precision matrix Σ^{-1}. A popular approach for obtaining a permutation invariant estimator which is sparse in the estimated precision matrix $\widehat{\Sigma}^{-1}$ is given by the ℓ_1-norm regularized maximum-likelihood estimation, also known as the GLasso Yuan and Lin (2007); Friedman et al. (2007); Banerjee et al. (2008). The GLasso approach is simple to use, at least when relying on publicly available software such as the `glasso` package in `R`. Further improvements have been reported when using some SCAD-type penalized maximum-likelihood estimator Lam and Fan (2009) or an adaptive GLasso procedure Fan et al. (2009), which can be thought of as a two-stage procedure. It is well-known from linear regression that such two- or multi-stage methods effectively address some bias problems which arise from ℓ_1-penalization Zou (2006); Candès and Tao (2007); Meinshausen (2007); Zou and Li (2008); Bühlmann and Meier (2008); Zhou (2009, 2010b).

In this paper we develop a new method for estimating graphical structure and parameters for multivariate Gaussian distributions using a multi-step procedure, which we call G**e**lat**o** (Graph **e**stimation with Lasso and Thresh**o**lding). Based on an ℓ_1-norm regularization and thresholding method in a first stage, we infer a sparse undirected graphical model, i.e. an estimated Gaussian conditional independence graph, and we then perform unpenalized maximum likelihood estimation (MLE) for the covariance Σ and its inverse Σ^{-1} based on the estimated graph. We make the following theoretical contributions: (i) Our method allows us to select a graphical structure which is sparse. In some sense we select only the important edges even though there may be many non-zero edges in the graph. (ii) Secondly, we evaluate the quality of the graph we have selected by showing consistency and establishing a rate of convergence in Frobenius norm of the estimated inverse covariance matrix; under sparsity constraints, the latter is of lower order than the corresponding results for the GLasso Rothman et al. (2008) and for the SCAD-type estimator Lam and Fan (2009). (iii) We show predictive risk consistency and provide a rate of convergence of the estimated covariance matrix. (iv) Lastly, we show general results for the MLE, where only *approximate* graph structures are given as input. Here, we explicitly analyze the performance of the maximum likelihood estimator as defined in (13) in all three metrics as just mentioned. Besides these theoretical advantages, we found empirically that our graph based method performs better in general, and sometimes substantially better than the GLasso, while we never found it clearly worse. Finally, our algorithm is simple and is comparable to the GLasso both in terms of computational time and implementation complexity.

There are a few key motivations and consequences for proposing such an approach based on graphical modeling. We will theoretically show that there are cases where our graph based method can accurately estimate conditional independencies among variables, i.e. the zeroes of Σ^{-1}, in situations where GLasso fails. The fact that GLasso easily fails to estimate the zeroes of Σ^{-1} has been recognized by Meinshausen (2008) and it has been discussed in more details in Ravikumar et al. (2008). Closer relations to existing work are primarily regarding our first stage of estimating the structure of the graph. We follow the nodewise regression approach from Meinshausen and Bühlmann (2006) but we make use of recent results for variable selection in linear models assuming the much weaker restricted eigenvalue condition Bickel et al. (2009); Zhou (2010b) instead of the restrictive neighborhood stability condition Meinshausen and Bühlmann (2006) or the equivalent irrepresentable condition Zhao and Yu (2006). In some sense, the novelty of our theory extending beyond Zhou (2010b) is the analysis for covariance and inverse covariance estimation and for risk consistency based on an estimated sparse graph as we mentioned above. Our regression and thresholding results build upon analysis of the thresholded Lasso estimator as studied in Zhou (2010b). Throughout our analysis, the sample complexity is one of the key focus point, which builds upon results in Zhou (2010a). Once the zeros are found, a constrained maximum likelihood estimator of the covariance can be computed, which was shown in Chaudhuri et al. (2007); it was unclear what the properties of such a procedure would be. Our theory answers such questions. As a two-stage method, our approach is also related to the adaptive Lasso Zou (2006) which has been analyzed for high-dimensional scenarios in Huang et al. (2008); Zhou et al. (2009); van de Geer et al. (2010). Another relation can be made to the method by Rütimann and Bühlmann (2009) for covariance and inverse covariance estimation based on a directed acyclic graph. This relation has only methodological character: the techniques and algorithms used in Rütimann and Bühlmann (2009) are very different and from a practical point of view, their approach has much higher degree of complexity in terms of computation and implementation, since estimation of an equivalence class of directed acyclic graphs is difficult and cumbersome.

Notation. We use the following notation. Given a graph $G = (V, E_0)$, where $V = \{1, \ldots, p\}$ is the set of vertices and E_0 is the set of undirected edges. we use s^i to denote the degree for node i, that is, the number of edges in E_0 connecting to node i. For an edge set E, we let $|E|$ denote its size. We use $\Theta_0 = \Sigma_0^{-1}$ and Σ_0 to refer to the true precision and covariance matrices respectively from now on. We denote the number of non-zero elements of Θ by $\text{supp}\,(\Theta)$. For any matrix $W = (w_{ij})$, let $|W|$ denote the determinant of W, $\text{tr}(W)$ the trace of W. Let $\varphi_{\max}(W)$ and $\varphi_{\min}(W)$ be the largest and smallest eigenvalues, respectively. We write $\text{diag}(W)$ for a diagonal matrix with the same diagonal as W. The matrix Frobenius norm is given by $\|W\|_F = \sqrt{\sum_i \sum_j w_{ij}^2}$. The operator norm $\|W\|_2^2$ is given by $\varphi_{\max}(WW^T)$. We write $|\cdot|_1$ for the ℓ_1 norm of a matrix vectorized, i.e., for a matrix $|W|_1 = \|\text{vec}W\|_1 = \sum_i \sum_j |w_{ij}|$, and sometimes write $\|W\|_0$ for the number of non-zero entries in the matrix. For an index set T and a matrix $W = [w_{ij}]$, write $W_T \equiv (w_{ij} I((i,j) \in T))$, where $I(\cdot)$ is the indicator function.

2. The model and the method

We assume a multivariate Gaussian model

$$X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma_0), \quad \text{where } \Sigma_{0,ii} = 1. \tag{1}$$

The data is generated by $X^{(1)}, \ldots, X^{(n)}$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma_0)$. Requiring the mean vector and all variances being equal to zero and one respectively is not a real restriction and in practice, we can easily center and scale the data. We denote the concentration matrix by $\Theta_0 = \Sigma_0^{-1}$.

Since we will use a nodewise regression procedure, as described below in Section 2.1, we consider a regression formulation of the model. Consider many regressions, where we regress one variable against all others:

$$X_i = \sum_{j \neq i} \beta_j^i X_j + V_i \ (i = 1, \ldots, p), \quad \text{where} \tag{2}$$

$$V_i \sim \mathcal{N}(0, \sigma^2_{V_i}) \text{ independent of } \{X_j; j \neq i\} \ (i = 1, \ldots, p). \tag{3}$$

There are explicit relations between the regression coefficients, error variances and the concentration matrix $\Theta_0 = (\theta_{0,ij})$:

$$\beta_j^i = -\theta_{0,ij}/\theta_{0,ii}, \ \mathrm{Var}(V_i) := \sigma^2_{V_i} = 1/\theta_{0,ii} \ (i, j = 1, \ldots, p). \tag{4}$$

Furthermore, it is well known that for Gaussian distributions, conditional independence is encoded in Θ_0, and due to (4), also in the regression coefficients:

$$\begin{aligned} & X_i \text{ is conditionally dependent of } X_j \text{ given } \{X_k;\ k \in \{1, \ldots, p\} \setminus \{i, j\}\} \\ \iff \quad & \theta_{0,ij} \neq 0 \quad \iff \quad \beta_i^j \neq 0 \text{ and } \beta_j^i \neq 0. \end{aligned} \tag{5}$$

For the second equivalence, we assume that $\mathrm{Var}(V_i) = 1/\theta_{0,ii} > 0$ and $\mathrm{Var}(V_j) = 1/\theta_{0,jj} > 0$. Conditional (in-)dependencies can be conveniently encoded by an undirected graph, the conditional independence graph which we denote by $G = (V, E_0)$. The set of vertices is $V = \{1, \ldots, p\}$ and the set of undirected edges $E_0 \subseteq V \times V$ is defined as follows:

$$\begin{aligned} & \text{there is an undirected edge between nodes } i \text{ and } j \\ \iff \quad & \theta_{0,ij} \neq 0 \quad \iff \quad \beta_i^j \neq 0 \text{ and } \beta_j^i \neq 0. \end{aligned} \tag{6}$$

Note that on the right hand side of the second equivalence, we could replace the word ”and” by ”or”. For the second equivalence, we assume $\mathrm{Var}(V_i), \mathrm{Var}(V_j) > 0$ following the remark after (5).

We now define the sparsity of the concentration matrix Θ_0 or the conditional independence graph. The definition is different than simply counting the non-zero elements of Θ_0, for which we have $\mathrm{supp}\,(\Theta_0) = p + 2|E_0|$. We consider instead the number of elements which are sufficiently large. For each i, define the number $s^i_{0,n}$ as the smallest integer such that the following holds:

$$\sum_{j=1, j \neq i}^{p} \min\{\theta^2_{0,ij}, \lambda^2 \theta_{0,ii}\} \ \leq \ s^i_{0,n} \lambda^2 \theta_{0,ii}, \ \text{ where } \lambda = \sqrt{2 \log(p)/n}, \tag{7}$$

where *essential sparsity* $s^i_{0,n}$ at row i describes the number of "sufficiently large" non-diagonal elements $\theta_{0,ij}$ relative to a given (n, p) pair and $\theta_{0,ii}, i = 1, \ldots, p$. The value $S_{0,n}$ in (8) is summing *essential sparsity* across all rows of Θ_0,

$$S_{0,n} \quad := \quad \sum_{i=1}^{p} s^i_{0,n}. \tag{8}$$

Due to the expression of λ, the value of $S_{0,n}$ depends on p and n. For example, if all non-zero non-diagonal elements $\theta_{0,ij}$ of the ith row are larger in absolute value than $\lambda\sqrt{\theta_{0,ii}}$, the value $s^i_{0,n}$ coincides with the node degree s^i. However, if some (many) of the elements $|\theta_{0,ij}|$ are non-zero but small, $s^i_{0,n}$ is (much) smaller than its node degree s^i; As a consequence, if some (many) of $|\theta_{0,ij}|, \forall i, j, i \neq j$ are non-zero but small, the value of $S_{0,n}$ is also (much) smaller than $2|E_0|$, which is the "classical" sparsity for the matrix $(\Theta_0 - \mathrm{diag}(\Theta_0))$. See Section A for more discussions.

2.1 The estimation procedure

The estimation of Θ_0 and $\Sigma_0 = \Theta_0^{-1}$ is pursued in two stages. We first estimate the undirected graph with edge set E_0 as in (6) and we then use the maximum likelihood estimator based on the estimate $\widehat{E}_n$, that is, the non-zero elements of $\widehat{\Theta}_n$ correspond to the estimated edges in $\widehat{E}_n$. Inferring the edge set E_0 can be based the following approach as proposed and theoretically justified in Meinshausen and Bühlmann (2006): perform p regressions using the Lasso to obtain p vectors of regression coefficients $\widehat{\beta}^1, \ldots, \widehat{\beta}^p$ where for each i, $\widehat{\beta}^i = \{\widehat{\beta}^i_j;\ j \in \{1, \ldots, p\} \setminus i\}$; Then estimate the edge set by the "OR" rule,

$$\text{estimate an edge between nodes } i \text{ and } j \Longleftrightarrow \widehat{\beta}^i_j \neq 0 \text{ or } \widehat{\beta}^j_i \neq 0. \tag{9}$$

Nodewise regressions for inferring the graph. In the present work, we use the Lasso in combination with thresholding Zhou (2010b). Consider the Lasso for each of the nodewise regressions

$$\beta^i_{\text{init}} = \text{argmin}_{\beta^i} \sum_{r=1}^{n} (X^{(r)}_i - \sum_{j \neq i} \beta^i_j X^{(r)}_j)^2 + \lambda_n \sum_{j \neq i} |\beta^i_j| \quad \text{for } i = 1, \ldots, p, \tag{10}$$

where $\lambda_n > 0$ is the same regularization parameter for all regressions. Since the Lasso typically estimates too many components with non-zero estimated regression coefficients, we use thresholding to get rid of variables with small regression coefficients from solutions of (10):

$$\widehat{\beta}^i_j(\lambda_n, \tau) = \beta^i_{j,\text{init}}(\lambda_n) I(|\beta^i_{j,\text{init}}(\lambda_n)| > \tau), \tag{11}$$

where $\tau > 0$ is a thresholding parameter. We obtain the corresponding estimated edge set as defined by (9) using the estimator in (11) and we use the notation

$$\widehat{E}_n(\lambda_n, \tau). \tag{12}$$

We note that the estimator depends on two tuning parameters λ_n and τ.

Maximum likelihood estimation based on graphs. Given a conditional independence graph with edge set E, we estimate the concentration matrix by maximum likelihood:

$$\widehat{\Theta}_n(E) = \text{argmin}_{\Theta \in \mathcal{M}_{p,E}} \left(\text{tr}(\Theta \widehat{S}_n) - \log|\Theta| \right), \text{ where}$$
$$\mathcal{M}_{p,E} = \{\Theta \in \mathbb{R}^{p \times p};\ \Theta \succ 0 \ \text{ and } \theta_{0,ij} = 0 \text{ for all } (i,j) \notin E, \ \text{ where } \ i \neq j\} \tag{13}$$

defines the constrained set for positive definite Θ and $\widehat{S}_n = n^{-1}\sum_{r=1}^{n} X^{(r)}(X^{(r)})^T$ is the sample covariance estimator (using that the mean vector is zero). The estimator in (13) is the maximum likelihood estimator with constraints to zero-values corresponding to the non-edges E^c:

$$E^c \quad = \quad \{(i,j) : i,j = 1,\ldots,p, i \neq j, (i,j) \notin E\}. \tag{14}$$

If the edge set E is sparse having relatively few edges only, the estimator in (13) is already sufficiently regularized by the constraints and hence, no additional penalization is used at this stage. Our final estimator for the concentration matrix is the combination of (12) and (13):

$$\widehat{\Theta}_n = \widehat{\Theta}_n(\widehat{E}_n(\lambda_n, \tau)). \tag{15}$$

Choosing the regularization parameters. We propose to select the parameter λ_n via cross-validation to minimize the squared test set error among all p regressions:

$$\widehat{\lambda}_n = \text{argmin}_{\lambda} \sum_{i=1}^{p} \left(\text{CV-score}(\lambda) \text{ of } i\text{th regression}\right),$$

where CV-score(λ) of ith regression is with respect to the squared error prediction loss. Sequentially proceeding, we then select τ by cross-validating the multivariate Gaussian log-likelihood, from (13). Regarding the type of cross-validation, we usually use the 10-fold scheme. Due to the sequential nature of choosing the regularization parameters, the number of candidate estimators is given by the number of candidate values for λ plus the number of candidate value for τ. In Section 4, we describe the grids of candidate values in more details. We note that for our theoretical results, we do not analyze the implications of our method using estimated $\widehat{\lambda}_n$ and $\widehat{\tau}$.

3. Theoretical results

In this section, we present in Theorem 1 convergence rates for estimating the precision and the covariance matrices with respect to the Frobenius norm; in addition, we show a risk consistency result for an oracle risk to be defined in (17). More importantly, we show the model we select is sufficiently sparse while at the same time, the bias term we introduce via sparse approximation is sufficiently bounded as given explicitly in Proposition 2. These results again illustrate the classical bias and variance tradeoff. Our analyses are non-asymptotic in nature; however, we first formulate our results from an asymptotic point of view for simplicity. To do so, we consider a triangular array of data generating random variables

$$X^{(1)}, \ldots, X^{(n)} \text{ i.i.d.} \sim \mathcal{N}_p(0, \Sigma_0),\ n = 1, 2, \ldots \tag{16}$$

where $\Sigma_0 = \Sigma_{0,n}$ and $p = p_n$ change with n. We make the following assumptions. Let $\Theta_0 := \Sigma_0^{-1}$.

(A0) The size of the neighborhood for each node $i \in V$ is upper bounded by an integer $s < p/2$.

(A1) The dimension and number of sufficiently strong non-zero edges $S_{0,n}$ as in (8) satisfy: dimension p grows with n following $p \asymp n^c$ for some constant $0 < c < 1$ and

$$p + S_{0,n} = o(n/\log(n)) \ (n \to \infty).$$

(A2) The minimal and maximal eigenvalues of the true covariance matrix Σ_0 are bounded: for some constants $M_{\text{upp}} \geq M_{\text{low}} > 0$, we have

$$\varphi_{\min}(\Sigma_0) \geq M_{\text{low}} > 0 \ \text{ and } \ \varphi_{\max}(\Sigma_0) \leq M_{\text{upp}} \leq \infty.$$

Moreover, throughout our analysis, we assume the following. There exists $v^2 > 0$ such that for all i, and V_i as defined in (3): $\quad \text{Var}(V_i) = 1/\theta_{0,ii} \geq v^2$.

For more discussions on these conditions, see Section A. Before we proceed, we need some definitions. Define for $\Theta \succ 0$

$$R(\Theta) = \text{tr}(\Theta\Sigma_0) - \log|\Theta|, \tag{17}$$

where minimizing (17) without constraints gives Θ_0. Given (8), (7), and Θ_0, define

$$C^2_{\text{diag}} := \min\{\max_{i=1,\ldots p} \theta^2_{0,ii}, \max_{i=1,\ldots,p} \left(s^i_{0,n}/S_{0,n}\right) \cdot \|\text{diag}(\Theta_0)\|^2_F\}. \tag{18}$$

We now state the main results of this paper. We defer the specification on various tuning parameters, namely, λ_n, τ to Section 3.2.

Theorem 1 *Consider data generating random variables as in (16) and assume that (A0), (A1), and (A2) hold. Then, with probability at least $1 - d/p^2$, for some small constant $d > 2$, we obtain under appropriately chosen λ_n and τ, an edge set $\widehat{E}_n$ as in (12), such that*

$$|\widehat{E}_n| \leq 4S_{0,n}, \quad \textit{where} \quad |\widehat{E}_n \setminus E_0| \leq 2S_{0,n}; \tag{19}$$

and for $\widehat{\Theta}_n$ and $\widehat{\Sigma}_n = (\widehat{\Theta}_n)^{-1}$ as defined in (15) the following holds,

$$\begin{aligned}
\|\widehat{\Theta}_n - \Theta_0\|_F &= O_P\left(\sqrt{(p + S_{0,n})\log(n)/n}\right), \\
\|\widehat{\Sigma}_n - \Sigma_0\|_F &= O_P\left(\sqrt{(p + S_{0,n})\log(n)/n}\right), \\
R(\widehat{\Theta}_n) - R(\Theta_0) &= O_P\left((p + S_{0,n})\log(n)/n\right)
\end{aligned}$$

where the contants hidden in the $O_P()$ notation depend on τ, $M_{\text{low}}, M_{\text{upp}}$, C_{diag} as in (18), and constants concerning sparse and restrictive eigenvalues of Σ_0 (cf. Section 3.2 and B).

The predictive risk can be interpreted as follows. Let $X \sim \mathcal{N}(0, \Sigma_0)$ with f_{Σ_0} denoting its density. Let $f_{\widehat{\Sigma}_n}$ be the density for $\mathcal{N}(0, \widehat{\Sigma}_n)$ and $D_{\text{KL}}(\Sigma_0 \| \widehat{\Sigma}_n)$ denotes the Kullback Leibler (KL) divergence from $\mathcal{N}(0, \Sigma_0)$ to $\mathcal{N}(0, \widehat{\Sigma}_n)$. Now, we have for $\Sigma, \widehat{\Sigma}_n \succ 0$,

$$R(\widehat{\Theta}_n) - R(\Theta_0) := 2\mathbf{E}_0 \left[\log f_{\Sigma_0}(X) - \log f_{\widehat{\Sigma}_n}(X)\right] := 2D_{\text{KL}}(\Sigma_0 \| \widehat{\Sigma}_n) \geq 0.$$

In Section 3.2, we provide an outline for achieving Theorem 1. The conditions that we use are indeed similar to those in Rothman et al. (2008), with (A1) being much more relaxed when $S_{0,n} \ll |E_0|$. We note that the bounded neighborhood constraint (A0) is required only for regression analysis (cf. Theorem 10) and for bounding the bias due to sparse approximation as in Proposition 2. We believe it can be relaxed when we do not aim to recover the graph structure. See Zhou (2010b) for more discussions on this point. Actual conditions and non-asymptotic results that are involved in the G**e**lat**o** estimation appear in Sections B, C, and D respectively.

Theorem 1 can be interpreted as follows. First, the cardinality of the estimated edge set exceeds $S_{0,n}$ at most by a factor 4, where $S_{0,n}$ as in (8) is the number of sufficiently strong edges in the model, while the number of false positives is bounded by $2S_{0,n}$. Note that the factors 4 and 2 can be replaced by some other constants, while achieving the same bounds on Frobenius norm (cf. Section D.1). We emphasize that we achieve these two goals by sparse model selection, where only important edges are selected even though there are many more non-zero edges in E_0, under conditions that are in some sense much weaker than (A2); For example, (A2) can be replaced by conditions on sparse and restrictive eigenvalues of Σ_0, much in the setting of Candès and Tao (2007); Meinshausen and Yu (2009); Bickel et al. (2009) for estimating regression coefficients except that we now impose such conditions on Σ_0 instead of the (regression) design matrix. Second, for the Frobenius norm and the risk to converge to zero, a too large value of p is not allowed and hence, a real high-dimensional scenario where $p \gg n$ is excluded. Hence (A1) is brought in only for this purpose. However, this restriction comes from the nature of the Frobenius norm and when considering e.g. the operator norm, such restrictions typically can be relaxed, see Rothman et al. (2008). The convergence rate with respect to the Frobenius norm should be compared to the rate $O_P(\sqrt{(p + |E_0|)\log(n)/n})$ which is the rate in Rothman et al. (2008) for the GLasso and for SCAD Lam and Fan (2009). In the scenario where $|E_0| \gg S_{0,n}$, i.e. there are many weak edges, the rate in Theorem 1 is better than the one established for GLasso Rothman et al. (2008) or for the SCAD-type estimator Lam and Fan (2009); hence we require a smaller sample size in order to yield an accurate estimate of Θ_0. We note that convergence rates for the estimated covariance matrix and for predictive risk depend on the rate in Frobenius norm of the estimated inverse covariance matrix. Finally, it is also of interest to understand the bias of the estimator caused by using the estimated edge set $\widehat{E}_n$ instead of the true edge set E_0. This is the content of Proposition 2. For a given $\widehat{E}_n$, denote by

$$\widetilde{\Theta}_0 = \text{diag}(\Theta_0) + (\Theta_0)_{\widehat{E}_n} = \text{diag}(\Theta_0) + \Theta_{0,\widehat{E}_n \cap E_0},$$

where the second equality holds since $\Theta_{0,E_0^c} = 0$. Note that the quantity $\widetilde{\Theta}_0$ is identical to Θ_0 on $\widehat{E}_n$ and on the diagonal, and it equals zero on $\widehat{E}_n^c$ as in (14). Hence, the quantity $\Theta_{0,\mathcal{D}} := \widetilde{\Theta}_0 - \Theta_0$ measures the bias caused by a potentially wrong edge set $\widehat{E}_n$; note that $\widetilde{\Theta}_0 = \Theta_0$ if $\widehat{E}_n = E_0$.

Proposition 2 *Consider data generating random variables as in expression (16). Assume that (A0) and (A2) hold and that* $\max\{p, S_{0,n}\} = o(n/\log(p))(n \to \infty)$. *Then we have for choices on* λ_n, τ *as in Theorem 1 and* $\widehat{E}_n$ *in (12),*

$$\|\Theta_{0,\mathcal{D}}\|_F := \|\widetilde{\Theta}_0 - \Theta_0\|_F = O_P\left(\sqrt{S_{0,n}\log(p)/n}\right).$$

We note that we achieve essentially the same rate for $\|(\widetilde{\Theta}_0)^{-1} - \Sigma_0\|_F$; see Remark 22. We give an account on how results in Proposition 2 are obtained in Section 3.2, with its non-asymptotic statement appearing in Corollary 12. Note that the sample size of $n = \Omega\left(\max(p, S_{0,n})\log n\right)$ as in Proposition 2 is less stringent than that implicitly specified in (A1), where we have specified a lower bound on the sample size to be $n = \Omega\left((p + S_{0,n})\log n\right)$. As to be shown in our analysis, the lower bound on n is slightly different for each Frobenius norm bound to hold from a non-asymptotic point of view (cf. Theorem 14 and 15).

3.1 Discussions and connections to previous work

It is interesting that the accuracy in terms of $\left\|\widehat{\Theta}_n - \Theta_0\right\|_F$ is not depending too strongly on the property to recover the true underlying edge set E_0 using (13). Regarding the latter, suppose we obtain with high probability the screening property

$$E \supseteq E_0, \tag{20}$$

when assuming that all non-zero regression coefficients $|\beta^i_j|$ are sufficiently large (E might be an estimate and hence random). Although we do not intend to make precise the exact conditions and choices of tuning parameters in regression and thresholding in order to achieve (20), we state Theorem 3, in case (20) holds with the following condition: the number of false positives is bounded as $|E \setminus E_0| \asymp p + S$. For simplicity, we state an asymptotic bound on the rate of convergence in Frobenius norm of the estimated $\widehat{\Theta}_n$.

Theorem 3 *Consider data generating random variables as in expression (16) and assume that (A1) and (A2) hold, where we replace* $S_{0,n}$ *with* $S := |E_0| = \sum_{i=1}^p s^i$. *Suppose on some event* $\mathcal{E}$, *such that* $\mathbb{P}(\mathcal{E}) \geq 1 - d/p^2$ *for a small constant* d, *we obtain an edge set* E *such that (20) holds and* $|E \setminus E_0| = O(S + p)$. *Let* $\widehat{\Theta}_n(E)$ *be the minimizer as defined in (13). Then, we have* $\|\widehat{\Theta}_n(E) - \Theta_0\|_F = O_P\left(\sqrt{(p + S)\log(n)/n}\right)$.

It is clear that this bound corresponds to exactly that of Rothman et al. (2008) for the GLasso estimation under appropriate choice of the penalty parameter. We omit the proof as it is more or less a simplified version of Theorem 14, which proves the stronger bounds as stated in Theorem 1, when E satisfies the sparsity conditions as in Theorem 1 and the bias condition in Proposition 2. We note that the maximum node-degree bound in (A0) is not needed for Theorem 3, nor for Theorem 14-16 to hold. We now make some connections to previous work. First, we note that to obtain with high probability the exact edge recovery, $E = E_0$, we need again sufficiently large non-zero edge weights and some restricted eigenvalue conditions on the covariance matrix as defined in Section A even for the multi-stage procedure. An earlier example is shown in Zhou et al. (2009), where the second stage estimator $\widehat{\beta}$ corresponding to (11) is obtained with nodewise regressions using adaptive Lasso Zou (2006) rather than thresholding as in the present work in order to recover the edge set E_0 with high probability. Clearly, given an accurate $\widehat{E}_n$, under (A1) and (A2) one can then apply Theorem 3 to accurately estimate $\widehat{\Theta}_n$. On the other hand, it is known that GLasso necessarily needs more restrictive conditions on Σ_0 than the nodewise regression approach with the Lasso, as discussed in Meinshausen (2007) and Ravikumar et al. (2008).

Furthermore, we believe it is easy to show that the nodewise regression approach with Lasso and thresholding (Gelato) works under the less restrictive assumptions on Σ_0 and with a smaller sample size than the analogue without the thresholding operation in order to achieve *nearly exact recovery* of the support in the sense that $\widehat{E}_n \supseteq E_0$ and $\max_i |\widehat{E}_{n,i} \setminus E_{0,i}|$ is small, which is to be understood as: the number of extra estimated edges at each node i is bounded by a small constant even when node degree s^i grows sublinearly with n for each i. This is shown essentially in Zhou (2010b) for single regression, in view of Theorem 25 in the present work. Given such properties of $\widehat{E}_n$, we can again apply Theorem 3 to obtain $\widehat{\Theta}_n$ under (A1) and (A2). In comparison to GLasso, Gelato requires weaker assumptions on Σ_0 in order to achieve the best sparsity and bias tradeoff as illustrated in Theorem 1 and Proposition 2 when many signals are weak, and Theorem 3 when all signals in E_0 are strong.

3.2 An outline for Theorem 1

Let $s_0 = \max_{i=1,\ldots,p} s^i_{0,n}$. We note that although sparse eigenvalues $\rho_{\max}(s), \rho_{\max}(3s_0)$ and restricted eigenvalue for Σ_0 (cf. Section A) are parameters that are unknown, we only need them to appear in the lower bounds for d_0, D_4, and hence also that for λ_n and t_0 that appear below. We simplify our notation in this section to keep it consistent with our theoretical non-asymptotic analysis to appear toward the end of this paper.

Regression. We choose for some $c_0 \geq 4\sqrt{2}$, $0 < \theta < 1$, and $\lambda = \sqrt{\log(p)/n}$,

$$\lambda_n = d_0 \lambda, \quad \text{where} \;\; d_0 \geq c_0 (1+\theta)^2 \sqrt{\rho_{\max}(s)\rho_{\max}(3s_0)}.$$

Let $\beta^i_{\text{init}}, i = 1, \ldots, p$ be the optimal solutions to (10) with λ_n as chosen above. We first prove an oracle result on nodewise regressions from Section 2.1 in Theorem 10.

Thresholding. We choose for some constants D_1, D_4 to be defined in Theorem 10,

$$t_0 = f_0 \lambda := D_4 d_0 \lambda \quad \text{where } D_4 \geq D_1$$

where D_1 depends on restrictive eigenvalue of Σ_0; Apply (11) with $\tau = t_0$ and $\beta^i_{\text{init}}, i = 1, \ldots, p$ for thresholding our initial regression coefficients. Let

$$\mathcal{D}^i = \{j : j \neq i, \; \left|\beta^i_{j,\text{init}}\right| < t_0 = f_0 \lambda\},$$

where bounds on $\mathcal{D}^i, i = 1, \ldots, p$ are given in Lemma 11. In view of (9), we let

$$\mathcal{D} = \{(i,j) : i \neq j : (i,j) \in \mathcal{D}^i \cap \mathcal{D}^j\}. \tag{21}$$

Selecting edge set E. Recall for a pair (i,j) we take the *OR rule* as in (9) to decide if it is to be included in the edge set E: for $\mathcal{D}$ as defined in (21), define

$$E := \{(i,j) : i,j = 1, \ldots, p, i \neq j, (i,j) \notin \mathcal{D}\}. \tag{22}$$

to be the subset of pairs of non-identical vertices of G which do not appear in $\mathcal{D}$; Let

$$\widetilde{\Theta}_0 = \mathrm{diag}(\Theta_0) + \Theta_{0,E_0 \cap E} \tag{23}$$

for E as in (22), which is identical to Θ_0 on all diagonal entries and entries indexed by $E_0 \cap E$, with the rest being set to zero. As shown in the proof of Corollary 12, by thresholding, we have identified a *sparse subset* of edges E of size at most $4S_{0,n}$, such that the corresponding bias $\|\Theta_{0,\mathcal{D}}\|_F := \|\widetilde{\Theta}_0 - \Theta_0\|_F$ is relatively small, i.e., as bounded in Proposition 2.

Refitting. In view of Proposition 2, we aim to recover $\widetilde{\Theta}_0$ given a sparse subset E; toward this goal, we use (13) to obtain the final estimator $\widehat{\Theta}_n$ and $\widehat{\Sigma}_n = (\widehat{\Theta}_n)^{-1}$. We give a more detailed account of this procedure in Section D, with a focus on elaborating the bias and variance tradeoff. We show the rate of convergence in Frobenius norm for the estimated $\widehat{\Theta}_n$ and $\widehat{\Sigma}_n$ in Theorem 14 and 15, and the bound for Kullback Leibler divergence in Theorem 16 respectively.

3.3 Discussion on covariance estimation based on maximum likelihood

The maximum likelihood estimate minimizes over all $\Theta \succ 0$,

$$\widehat{R}_n(\Theta) = \mathrm{tr}(\Theta \widehat{S}_n) - \log|\Theta| \tag{24}$$

where $\widehat{S}_n$ is the sample covariance matrix. Minimizing $\widehat{R}_n(\Theta)$ without constraints gives $\widehat{\Sigma}_n = \widehat{S}_n$. We now would like to minimize (24) under the constraints that some pre-defined subset $\mathcal{D}$ of edges are set to zero. Then the follow relationships hold regarding $\widehat{\Theta}_n(E)$ defined in (13) and its inverse $\widehat{\Sigma}_n$, and $\widehat{S}_n$: for E as defined in (22),

$$\begin{aligned} \widehat{\Theta}_{n,ij} &= 0, \ \forall (i,j) \in \mathcal{D} \ \text{ and} \\ \widehat{\Sigma}_{n,ij} &= \widehat{S}_{n,ij}, \ \forall (i,j) \in E \cup \{(i,i), i = 1, \ldots, p\}. \end{aligned}$$

Hence the entries in the covariance matrix $\widehat{\Sigma}_n$ for the chosen set of edges in E and the diagonal entries are set to their corresponding values in $\widehat{S}_n$. Indeed, we can derive the above relationships via the Lagrange form, where we add Lagrange constants γ_{jk} for edges in $\mathcal{D}$,

$$\ell_C(\Theta) = \log|\Theta| - \mathrm{tr}(\widehat{S}_n \Theta) - \sum_{(j,k) \in \mathcal{D}} \gamma_{jk} \theta_{0,jk}. \tag{25}$$

Now the gradient equation of (25) is:

$$\Theta^{-1} - \widehat{S}_n - \Gamma = 0,$$

where Γ is a matrix of Lagrange parameters such that $\gamma_{jk} \neq 0$ for all $(j, k) \in \mathcal{D}$ and $\gamma_{jk} = 0$ otherwise. Throughout this paper, we assume that graph $G = (V, E_0)$ is connected. Otherwise, the MLE problem can potentially be decomposed into a number of independent problems, for which we solve independently for each connected component. This will be one of the directions for our future work.

4. Numerical results

We note that the notation in this section is necessarily different from the rest of the paper to make things simple. In this section we compare the empirical performance of our estimation method with the GLasso for simulated and real data. The GLasso is defined as:

$$\widehat{\Theta}_{\text{GLasso}} = \underset{\Theta \succ 0}{\text{argmin}}(\text{tr}(\widehat{S}_n\Theta) - \log|\Theta| + \rho \sum_{i<j} |\theta_{ij}|)$$

where $\widehat{S}_n$ is the empirical covariance matrix and the minimization is over positive definite matrices. For computation of the Gelato, we used the R-packages glmnet Friedman et al. (2010) and glasso Friedman et al. (2007).

4.1 Simulation study

In our simulation study, we look at three different models.

- An AR(1)-Block model. In this model the covariance matrix is block-diagonal with equal-sized AR(1)-blocks of the form $\Sigma_{Block} = \{r^{|i-j|}\}_{i,j}$.
- The random concentration matrix model considered in Rothman et al. (2008). In this model, the concentration matrix is $\Theta = B + \delta I$ where each off-diagonal entry in B is generated independently and equal to 0 or 0.5 with probability $1 - \pi$ or π, respectively. All diagonal entries of B are zero, and δ is chosen such that the condition number of Θ is p.
- The exponential decay model considered in Fan et al. (2009). In this model we consider a case where no element of the concentration matrix is exactly zero. The elements of Θ are given by $\theta_{ij} = \exp(-2|i-j|)$ equals essentially zero when the difference $|i-j|$ is large.

We compare the two estimators for each model with $p = 300$ and $n = 40, 80, 320$. For each model we sample data $X^{(1)}, \ldots, X^{(n)}$ i.i.d. $\sim \mathcal{N}(0, \Sigma)$. We use two different performance measures. The Frobenius norm of the estimation error $\|\widehat{\Sigma} - \Sigma\|_F$ and $\|\widehat{\Theta} - \Theta\|_F$, and the Kullback-Leibler divergence between $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \widehat{\Sigma})$:

$$2D_{\text{KL}}(\Sigma\|\widehat{\Sigma}) = \text{tr}\left(\Sigma\widehat{\Theta}\right) - \log|\Sigma\widehat{\Theta}| - p := R(\widehat{\Theta}) - R(\Sigma^{-1})$$

for R as defined in (17). For the two estimation methods we have various tuning parameters, namely λ, τ and ρ. Due to the computational complexity we specify the two parameters of our Gelato method sequentially. That is, we derive the optimal value of the penalty parameter λ by 10-fold cross-validation with respect to the test set squared error for all the nodewise regressions. After fixing $\lambda = \lambda_{CV}$ we obtain the optimal threshold τ again by 10-fold cross-validation but with respect to the negative Gaussian log-likelihood. For the parameter ρ of the GLasso estimator we also use a 10-fold cross-validation with respect to the negative Gaussian log-likelihood. The grids of candidate values for the cross-validations are given as follows:

$$\lambda_r = A_r\sqrt{\frac{\log p}{n}} \quad r = 1, \dots, 10 \quad \text{with} \quad \tau_r = 0.75 \cdot \lambda_r$$
$$\rho_r = B_r\sqrt{\frac{\log p}{n}} \quad r = 1, \dots, 10$$

where $A_r, B_r \in \{0.01, 0.05, 0.1, 0.3, 0.5, 1, 2, 4, 8, 16\}$.
The two different performance measures are evaluated for the estimators based on the sample $X^{(1)}, \dots, X^{(n)}$ with optimal tuning parameters λ, τ and ρ for each model from above. All results are based on 50 independent simulation runs.

4.1.1 THE AR(1)-BLOCK MODEL

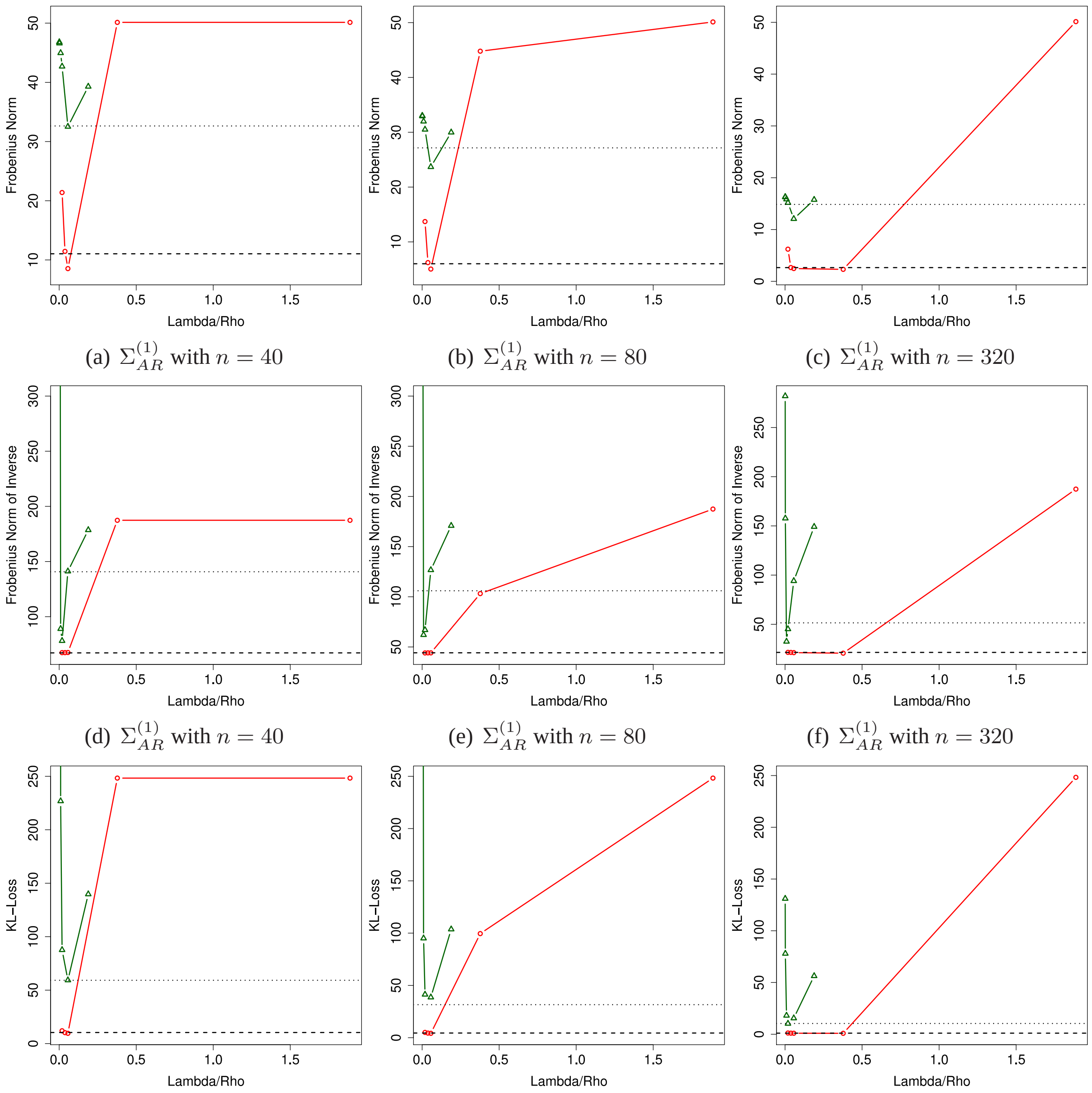
We consider two different covariance matrices. The first one is a simple auto-regressive process of order one with trivial block size equal to $p = 300$, denoted by $\Sigma_{AR}^{(1)}$. This is also known as a Toeplitz matrix. That is, we have $\Sigma_{AR;i,j}^{(1)} = r^{|i-j|}\ \forall\ i, j \in \{1, ..., p\}$. The second matrix $\Sigma_{AR}^{(2)}$ is a block-diagonal matrix with AR(1) blocks of equal block size 30×30, and hence the block-diagonal of $\Sigma_{AR}^{(2)}$ equals $\Sigma_{Block;i,j} = r^{|i-j|}$, $i, j \in \{1, \dots, 30\}$. For both models $\Sigma_{AR}^{(1)}$ and $\Sigma_{AR}^{(2)}$ we choose $r = 0.9$. The results of the simulation are shown in Figure 1 and 2.

The figures show a substantial performance gain of our method compared to the GLasso in both considered covariance models. This result speaks for our method, especially because AR(1)-block models are very simple.

4.1.2 THE RANDOM PRECISION MATRIX MODEL

For this model we also consider two different matrices, which differ in sparsity. For the sparser matrix $\Theta^{(3)}$ we set the probability π to 0.1. That is , we have an off diagonal entry in $\Theta^{(3)}$ of 0.5 with probability $\pi = 0.1$ and an entry of 0 with probability 0.9. In the case of the second matrix $\Theta^{(4)}$ we set π to 0.5 which provides us with a denser concentration matrix. The simulation results for the two performance measures are given in Figure 3 and 4.

From Figures 3 and 4 we see that Gelato keeps up with the GLasso in both the sparse and the dense simulation settings. It performs better than the GLasso with respect to $\|\widehat{\Theta} - \Theta\|_F$ and the Kullback Leibler divergence but is inferior for the Frobenius norm of $\widehat{\Sigma} - \Sigma$.



(a) $\Sigma_{AR}^{(1)}$ with $n = 40$

(b) $\Sigma_{AR}^{(1)}$ with $n = 80$

(c) $\Sigma_{AR}^{(1)}$ with $n = 320$

(d) $\Sigma_{AR}^{(1)}$ with $n = 40$

(e) $\Sigma_{AR}^{(1)}$ with $n = 80$

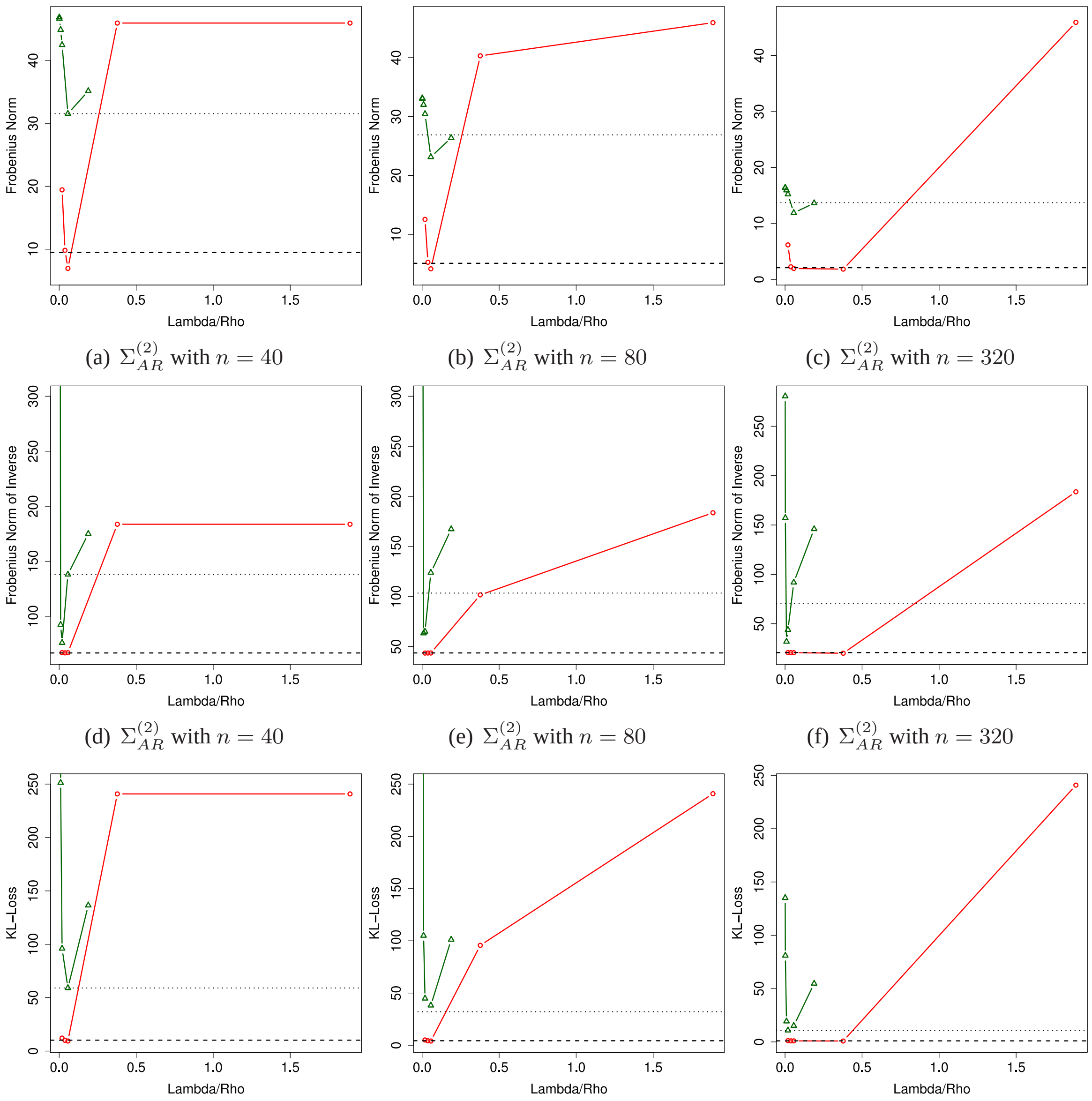
(f) $\Sigma_{AR}^{(1)}$ with $n = 320$

(g) $\Sigma_{AR}^{(1)}$ with $n = 40$

(h) $\Sigma_{AR}^{(1)}$ with $n = 80$

(i) $\Sigma_{AR}^{(1)}$ with $n = 320$

Figure 1: Plots for $\Sigma_{AR}^{(1)}$. The triangles (green) stand for the GLasso and the circles (red) for our Gelato method with a representative value of τ. The horizontal lines show the performances of the two techniques for cross-validated tuning parameters λ, τ and ρ. The dashed line stands for our Gelato method and the dotted one for the GLasso. Lambda/Rho stands for λ or ρ, respectively.



(a) $\Sigma_{AR}^{(2)}$ with $n = 40$

(b) $\Sigma_{AR}^{(2)}$ with $n = 80$

(c) $\Sigma_{AR}^{(2)}$ with $n = 320$

(d) $\Sigma_{AR}^{(2)}$ with $n = 40$

(e) $\Sigma_{AR}^{(2)}$ with $n = 80$

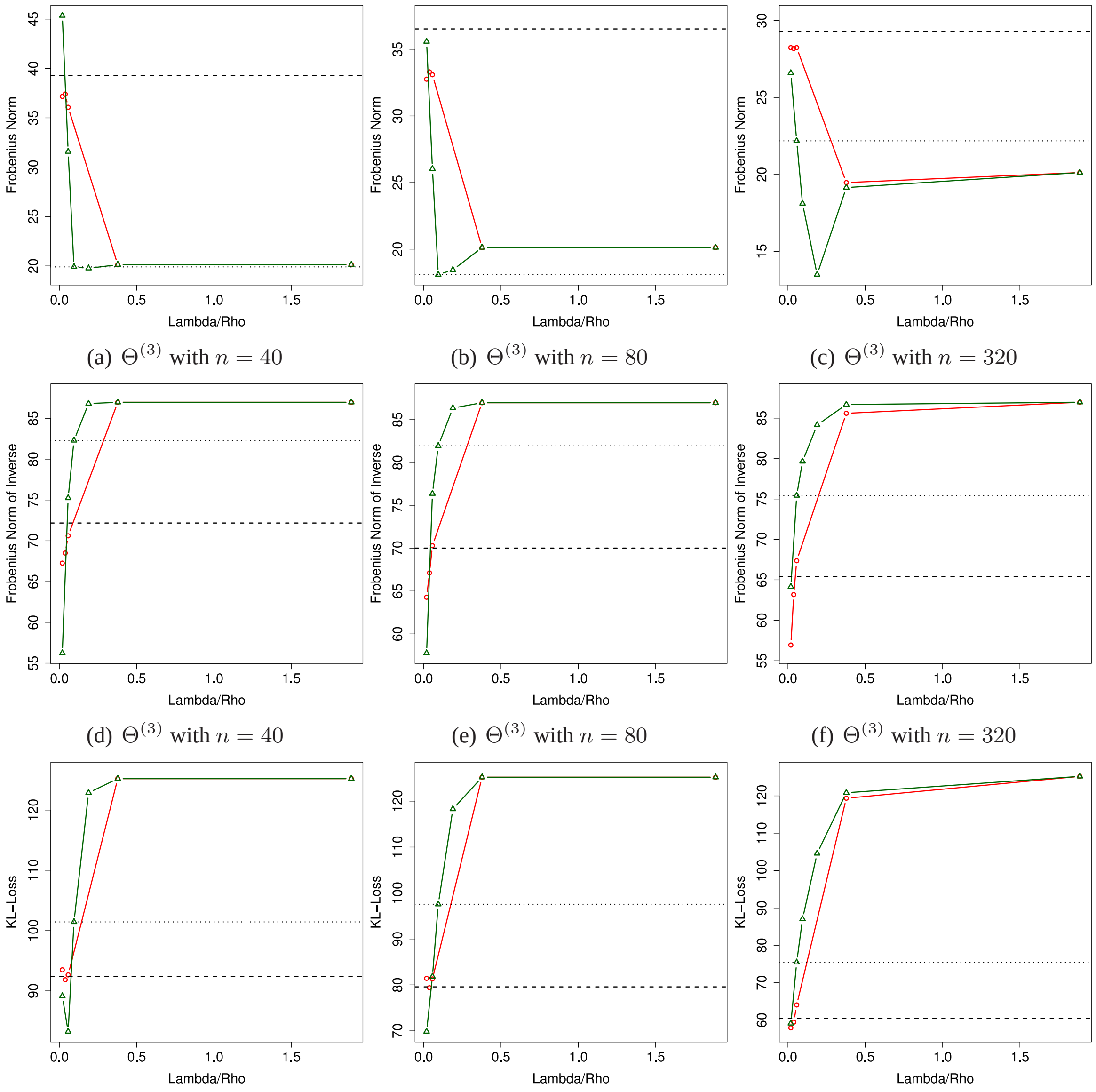
(f) $\Sigma_{AR}^{(2)}$ with $n = 320$

(g) $\Sigma_{AR}^{(2)}$ with $n = 40$

(h) $\Sigma_{AR}^{(2)}$ with $n = 80$

(i) $\Sigma_{AR}^{(2)}$ with $n = 320$

Figure 2: Plots for $\Sigma_{AR}^{(2)}$. The triangles (green) stand for the GLasso and the circles (red) for our Gelato method with a representative value of τ. The horizontal lines show the performances of the two techniques for cross-validated tuning parameters λ, τ and ρ. The dashed line stands for our Gelato method and the dotted one for the GLasso. Lambda/Rho stands for λ or ρ, respectively.



(a) $\Theta^{(3)}$ with $n = 40$

(b) $\Theta^{(3)}$ with $n = 80$

(c) $\Theta^{(3)}$ with $n = 320$

(d) $\Theta^{(3)}$ with $n = 40$

(e) $\Theta^{(3)}$ with $n = 80$

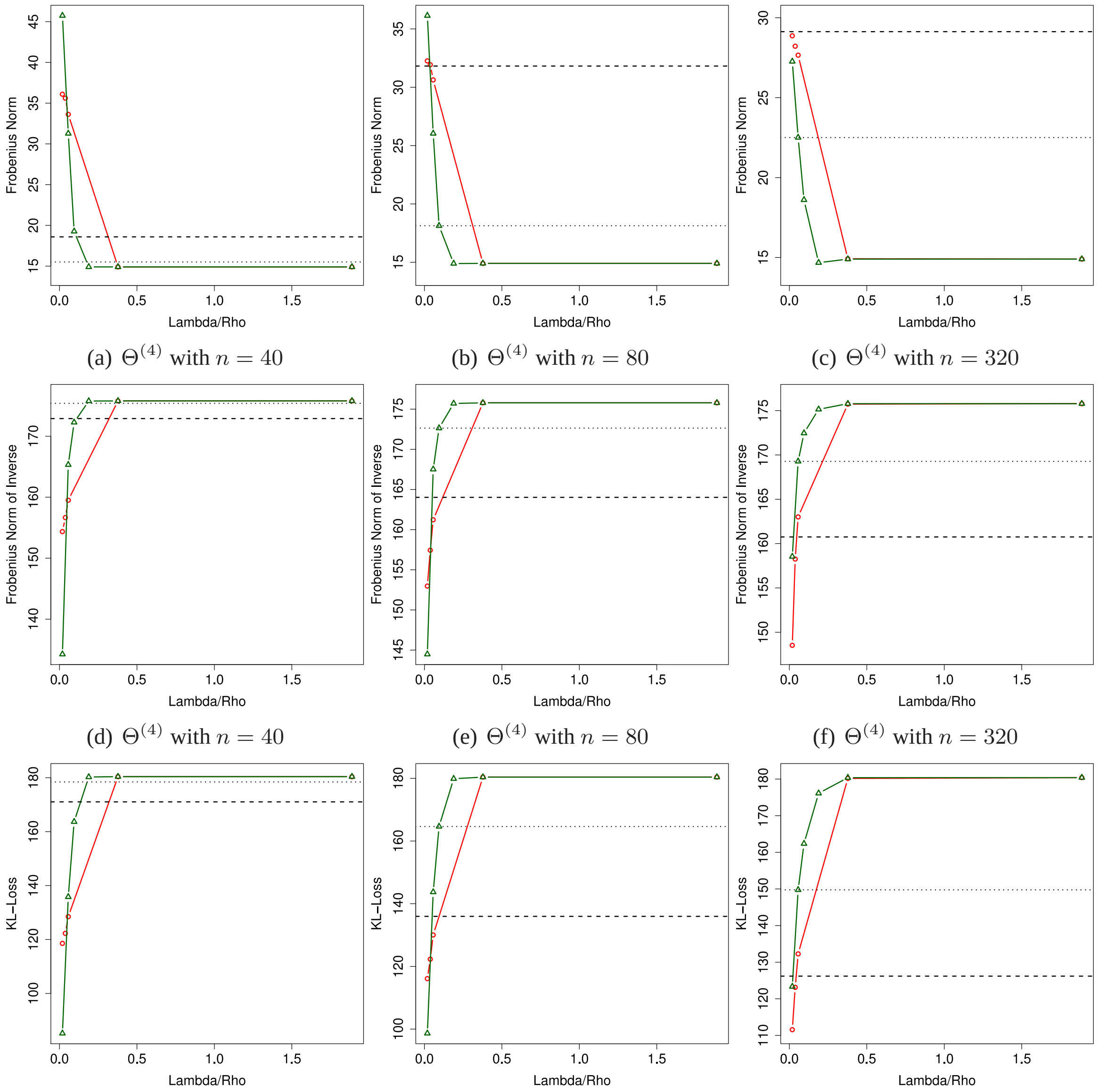
(f) $\Theta^{(3)}$ with $n = 320$

(g) $\Theta^{(3)}$ with $n = 40$

(h) $\Theta^{(3)}$ with $n = 80$

(i) $\Theta^{(3)}$ with $n = 320$

Figure 3: Plots for $\Theta^{(3)}$. The triangles (green) stand for the GLasso and the circles (red) for our Gelato method with a representative value of τ. The horizontal lines show the performances of the two techniques for cross-validated tuning parameters λ, τ and ρ. The dashed line stands for our Gelato method and the dotted one for the GLasso. Lambda/Rho stands for λ or ρ, respectively.



(a) $\Theta^{(4)}$ with $n = 40$

(b) $\Theta^{(4)}$ with $n = 80$

(c) $\Theta^{(4)}$ with $n = 320$

(d) $\Theta^{(4)}$ with $n = 40$

(e) $\Theta^{(4)}$ with $n = 80$

(f) $\Theta^{(4)}$ with $n = 320$

(g) $\Theta^{(4)}$ with $n = 40$

(h) $\Theta^{(4)}$ with $n = 80$

(i) $\Theta^{(4)}$ with $n = 320$

Figure 4: Plots for $\Theta^{(4)}$. The triangles (green) stand for the GLasso and the circles (red) for our Gelato method with a representative value of τ. The horizontal lines show the performances of the two techniques for cross-validated tuning parameters λ, τ and ρ. The dashed line stands for our Gelato method and the dotted one for the GLasso. Lambda/Rho stands for λ or ρ, respectively.

4.1.3 The exponential decay model

In this simulation setting we only have one version of the concentration matrix $\Theta^{(5)}$. The entries of $\Theta^{(5)}$ are generated by $\Theta^{(5)}_{i,j} = \exp(-2|i-j|)$.

Figure 5 shows the results of the simulation. We find that both methods show equal performances in both the Frobenius norm and the Kullback Leibler divergence. This is not entirely surprising as we expect Gelato to work best when Θ is relatively sparse.

4.2 Application to real data

4.2.1 Isoprenoid gene pathway in Arabidobsis thaliana

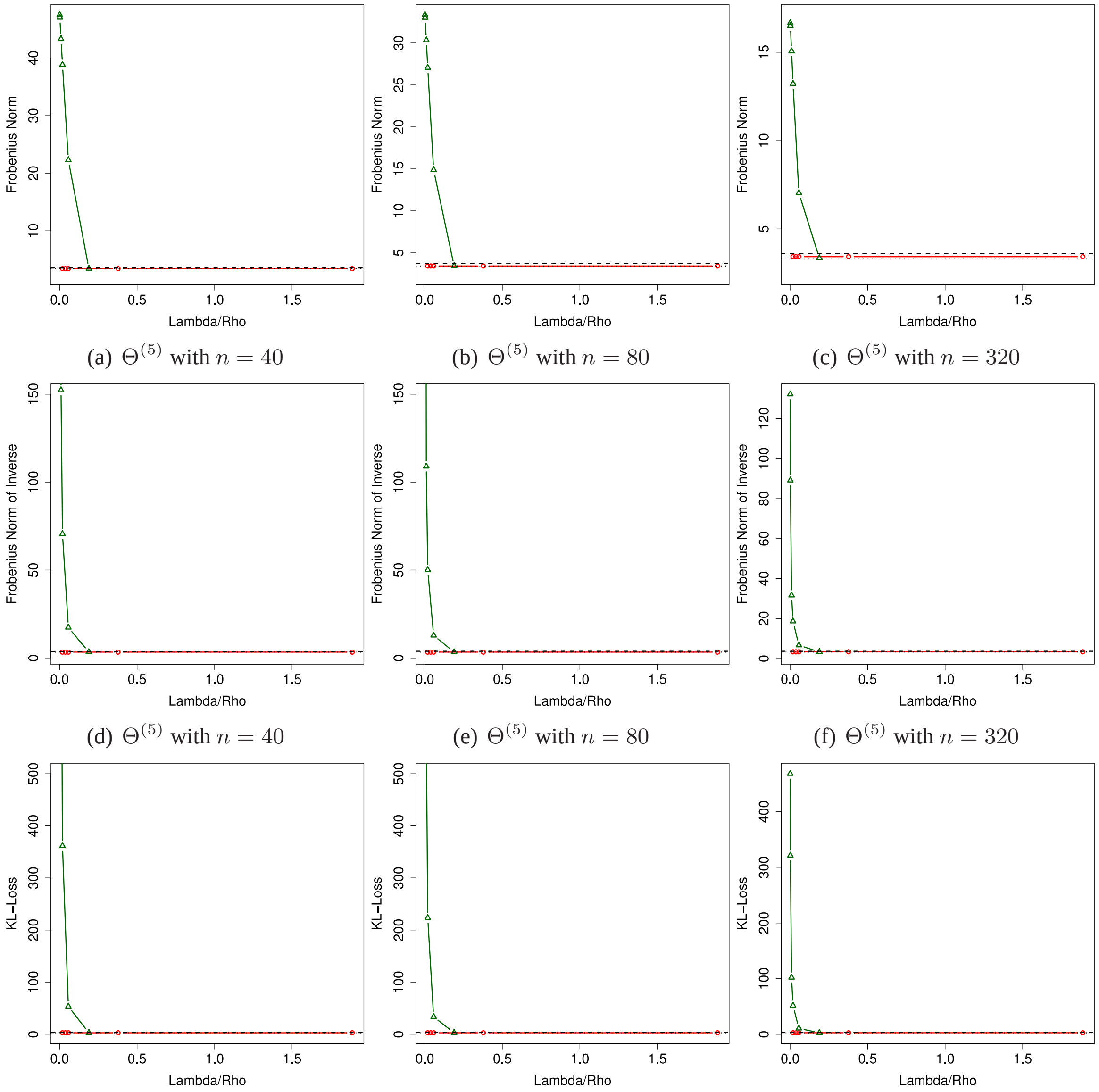
In this example we compare the two estimators on the isoprenoid biosynthesis pathway data given in Wille et al. (2004). Isoprenoids play various roles in plant and animal physiological processes and as intermediates in the biological synthesis of other important molecules. In plants they serve numerous biochemical functions in processes such as photosynthesis, regulation of growth and development.

The data set consists of $p = 39$ isoprenoid genes for which we have $n = 118$ gene expression patterns under various experimental conditions. In order to compare the two techniques we compute the negative log-likelihood via 10-fold cross-validation for different values of λ, τ and ρ. In Figure 6 we plot the cross-validated negative log-likelihood against the logarithm of the average number of non-zero entries (logarithm of the ℓ_0-norm) of the estimated concentration matrix $\widehat{\Theta}$. The logarithm of the ℓ_0-norm reflects the sparsity of the matrix $\widehat{\Theta}$ and therefore the figures show the performance of the estimators for different levels of sparsity. The plots do not allow for a clear conclusion. The GLasso performs slightly better when allowing for a rather dense fit. On the other hand, when requiring a sparse fit, the Gelato performs better.

4.2.2 Clinical status of human breast cancer

As a second example, we compare the two methods on the breast cancer dataset from West et al. (2001). The tumor samples were selected from the Duke Breast Cancer SPORE tissue bank. The data consists of $p = 7129$ genes with $n = 49$ breast tumor samples. For the analysis we use the 100 variables with the largest sample variance. As before, we compute the negative log-likelihood via 10-fold cross-validation. Figure 6 shows the results.

In this real data example the interpretation of the plots is similar as for the arabidopsis dataset. For dense fits, GLasso is better while Gelato has an advantage when requiring a sparse fit.



(a) $\Theta^{(5)}$ with $n = 40$

(b) $\Theta^{(5)}$ with $n = 80$

(c) $\Theta^{(5)}$ with $n = 320$

(d) $\Theta^{(5)}$ with $n = 40$

(e) $\Theta^{(5)}$ with $n = 80$

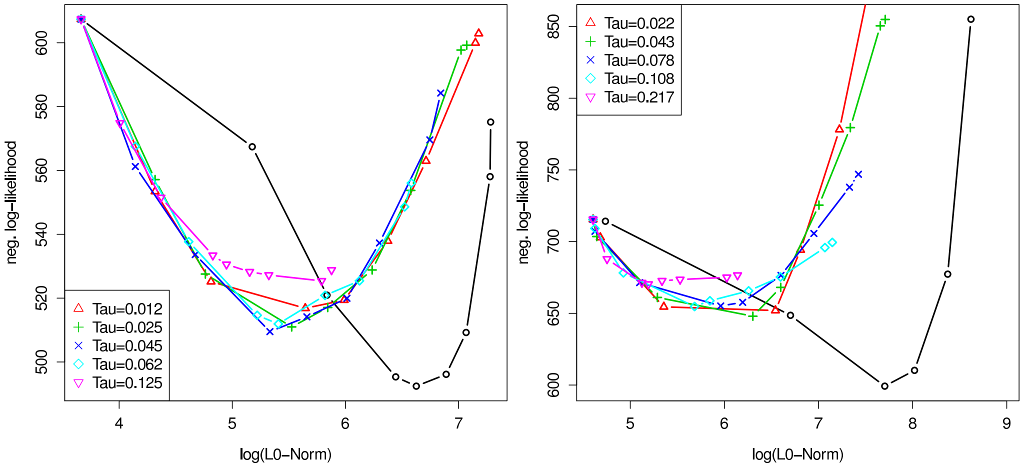
(f) $\Theta^{(5)}$ with $n = 320$

(g) $\Theta^{(5)}$ with $n = 40$

(h) $\Theta^{(5)}$ with $n = 80$

(i) $\Theta^{(5)}$ with $n = 320$

Figure 5: Plots for $\Theta^{(5)}$. The triangles (green) stand for the GLasso and the circles (red) for our Gelato method with a representative value of τ. The horizontal lines show the performances of the two techniques for cross-validated tuning parameters λ, τ and ρ. The dashed line stands for our Gelato method and the dotted one for the GLasso. Lambda/Rho stands for λ or ρ, respectively.



(a) isoprenoid data

(b) breast cancer data

Figure 6: Plots for the isporenoid data from arabidopsis thaliana (a) and the human breast cancer data (b). 10-fold cross-validation of negative log-likelihood against the logarithm of the average number of non-zero entries of the estimated concentration matrix $\widehat{\Theta}$. The circles stand for the GLasso and the Gelato is displayed for various values of τ.

5. Conclusions

We propose and analyze the Gelato estimator. Its advantage is that it automatically yields a positive definite covariance matrix and the Frobenius norm on its inverse has in some settings a better rate of convergence than the GLasso or SCAD type estimators. From a theoretical point of view, our method is clearly gauged for bounding the Frobenius norm of the inverse covariance matrix. We also derive bounds on the convergence rate for the estimated covariance matrix and on on the Kullback-Leibler divergence. From a non-asymptotic point of view, our method has a clear advantage when the sample size is small relative to the sparsity $S = |E_0|$: for a given sample size n, we bound the variance in our re-estimation stage by excluding edges of E_0 with small weights from the selected edge set $\widehat{E}_n$ while ensuring that we do not introduce too much bias. Our Gelato method also addresses the bias problem inherent in the GLasso estimator since we no longer shrink the entries in the covariance matrix corresponding to the selected edge set $\widehat{E}_n$ in the maximum likelihood estimate, as shown in Section 3.3.

Our experimental results show that when the graph is sparse, Gelato performs better (and sometimes substantially better, for example for AR(1)-Block models) than the GLasso consistently in all performance measures, and slightly worse only with respect to the Frobenius norm of the covariance matrix when the truth is a dense graph. We also show experimentally how one can use cross-validation for choosing the tuning parameters in regression and thresholding. Deriving theoretical results on cross-validation is not within the scope of this paper.

6. Acknowledgement

Min Xu's research was supported by NSF grant CCF-0625879 and AFOSR contract FA9550-09-1-0373. Shuheng Zhou thanks Bin Yu warmly for hosting her visit at UC Berkeley while she was conducting this research in Spring 2010. SZ's research was supported in part by the Swiss National Science Foundation (SNF) Grant 20PA21-120050/1. We thank conversations with Larry Wasserman, which inspired the name of G**e**lat**o**; see main text.

Appendix A. Theoretical analysis and proofs

In this section, we specify some preliminary definitions. First, note that when we discuss estimating the parameters Σ_0 and $\Theta_0 = \Sigma_0^{-1}$, we always assume that

$$\varphi_{\max}(\Sigma_0) := 1/\varphi_{\min}(\Theta_0) \leq 1/\underline{c} < \infty \text{ and } 1/\varphi_{\max}(\Theta_0) = \varphi_{\min}(\Sigma_0) \geq \underline{k} > 0, \quad (26)$$
$$\text{where we assume } \underline{k}, \underline{c} \leq 1 \quad \text{so that } \underline{c} \leq 1 \leq 1/\underline{k}. \quad (27)$$

It is clear that these conditions are exactly that of (A2) in Section 3 with

$$M_{\text{upp}} := 1/\underline{c} \quad \text{and} \quad M_{\text{low}} := \underline{k},$$

where it is clear that for $\Sigma_{0,ii} = 1, i = 1, \ldots, p$, we have the sum of p eigenvalues of Σ_0, $\sum_{i=1}^{p} \varphi_i(\Sigma_0) = \text{tr}(\Sigma_0) = p$. Hence it will make sense to assume that (27) holds, since otherwise, (26) implies that $\varphi_{\min}(\Sigma_0) = \varphi_{\max}(\Sigma_0) = 1$ which is unnecessarily restrictive.

We now define parameters relating to the key notion of *essential sparsity* s_0 as explored in Candès and Tao (2007); Zhou (2009, 2010b) for regression. Denote the number of non-zero non-diagonal entries in each row of Θ_0 by s^i. Let $s = \max_{i=1,\ldots,p} s^i$ denote the highest node degree in $G = (V, E_0)$. Consider nodewise regressions as in (2), where we are given vectors of parameters $\{\beta^i_j, j = 1, \ldots, p, j \neq i\}$ for $i = 1, \ldots, p$. With respect to the neighborhood of node i for each i, we define $s^i_0 \leq s^i \leq s$ as the smallest integer such that

$$\sum_{j=1, j\neq i}^{p} \min((\beta^i_j)^2, \lambda^2 \text{Var}(V_i)) \leq s^i_0 \lambda^2 \text{Var}(V_i), \text{ where } \lambda = \sqrt{2 \log p/n}. \quad (28)$$

Note that we drop subscript n from s^i_0, which coincides with $s^i_{0,n}$ as defined in (7). We use the following symbols as a shorthand throughout our proofs to simplify our notation.

Definition 4 (Bounded neighborhood parameters.) *The size of the neighborhood s^i for each node i is upper bounded by an integer $s < p/2$. For s^i_0 as in (28), define*

$$s_0 := \max_{i=1,\ldots,p} s^i_0 \leq s < p/2 \text{ and} \quad (29)$$
$$S_{0,n} := \sum_{i=1,\ldots,p} s^i_0 \leq s_0 p, \quad (30)$$

which coincides with $S_{0,n}$ as defined in (8).

Next, we define the following parameters that are relevant to nodewise regressions for a random design X. Recall the data is generated by $X^{(1)}, \ldots, X^{(n)}$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma_0)$, where $\Sigma_{0,ii} = 1$. For an integer $m < p$, we define **m-sparse eigenvalues** of Σ_0 as follows:

$$\sqrt{\rho_{\min}(m)} \quad := \quad \min_{\substack{t \neq 0 \\ |\operatorname{supp}(t)| \leq m}} \frac{\left\|\Sigma_0^{1/2} t\right\|_2}{\|t\|_2}, \quad \sqrt{\rho_{\max}(m)} \quad := \quad \max_{\substack{t \neq 0 \\ |\operatorname{supp}(t)| \leq m}} \frac{\left\|\Sigma_0^{1/2} t\right\|_2}{\|t\|_2}.$$

For a given sparsity parameter s_0 as defined in (29), we define the following condition, which was originally defined in Zhou et al. (2009), motivated by Bickel et al. (2009). It is clear that when s_0 and k_0 become smaller, $RE(s_0, k_0, \Sigma_0)$ condition becomes easier to hold with K becomes correspondingly smaller.

Definition 5 (Restricted eigenvalue condition $RE(s_0, k_0, \Sigma_0)$). *For some integer $1 \leq s_0 \leq p/2$ and a positive number k_0, the following condition holds for all $\upsilon \neq 0$,*

$$\frac{1}{K(s_0, k_0, \Sigma_0)} := \min_{\substack{J_0 \subseteq \{1, \ldots, p\}, \\ |J_0| \leq s_0}} \min_{\left\|\upsilon_{J_0^c}\right\|_1 \leq k_0 \left\|\upsilon_{J_0}\right\|_1} \frac{\left\|\Sigma_0^{1/2} \upsilon\right\|_2}{\|\upsilon_{J_0}\|_2} > 0, \tag{31}$$

where we assume $\Sigma_{0,jj} = 1, \forall j = 1, \ldots, p$.

In the context of Gaussian graphical modeling, where we only aim to estimate the graphical structure E_0 itself, (26) need not hold in general. Throughout the rest of the paper up till Section D, we assume that Σ_0 satisfies (31) for s_0 as in (28) and the sparse eigenvalue $\rho_{\min}(s) > 0$, where s is the maximum node degree in G; Clearly we have $\rho_{\max}(s) \leq s$ as we assume $\Sigma_{0,jj} = 1, \forall j = 1, \ldots, p$. In the context of covariance estimation, we do assume that (26) holds; in this case such RE condition always holds on Σ_0, and $\rho_{\max}(m), \rho_{\min}(m)$ are bounded by some constants for all $m \leq p$. In this case, we continue to adopt parameters such as K, $\rho_{\max}(s)$, and $\rho_{\max}(3s_0)$ for the purpose of defining constants that are reasonable tight under condition (26). In general, one can think of

$$\rho_{\max}(\max(3s_0, s)) \ll 1/\underline{c} < \infty \quad \text{and} \quad K^2(s_0, k_0, \Sigma_0) \ll 1/\underline{k} < \infty,$$

for $\underline{c}, \underline{k}$ as in (26) and s_0 as in (29).

Roughly speaking, for two variables X_i, X_j as in (1) such that their corresponding entry in $\Theta_0 = (\theta_{0,ij})$ satisfies: $\theta_{0,ij} < \lambda \sqrt{\theta_{0,ii}}$, where $\lambda = \sqrt{2 \log(p)/n}$, we can not guarantee that $(i, j) \in \widehat{E}_n$ when we aim to keep $\asymp s_0^i$ edges for node $i, i = 1, \ldots, p$. For a given Θ_0, it is clear that as sample size n increases, we are able to select edges with smaller coefficient $\theta_{0,ij}$. In fact it holds that

$$|\theta_{0,ij}| < \lambda \sqrt{\theta_{0,ii}} \text{ which is equivalent to } |\beta_j^i| < \lambda \sigma_{V_i}, \quad \text{for all } \; j \geq s_0^i + 1 + \mathbb{I}_{i \leq s_0^i + 1}, \tag{32}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function, if we order the regression coefficients as follows:

$$|\beta_1^i| \geq |\beta_2^i| ... \geq |\beta_{i-1}^i| \geq |\beta_{i+1}^i| .... \geq |\beta_p^i|,$$

in view of (2), which is the same as if we order for row i of Θ_0,

$$|\theta_{0,i1}| \geq |\theta_{0,i,2}| ... \geq |\theta_{0,i,i-1}| \geq |\theta_{0,i,i+1}| .... \geq |\theta_{0,i,p}|. \tag{33}$$

This has been show in (Candès and Tao, 2007); See also Zhou (2010b).

A.1 Concentration bounds for the random design

We assume $k_0 > 0$ and it is understood to be the same quantity throughout our discussion. In preparation for showing the oracle results of Lasso in Theorem 25, we first state some concentration bounds on the random design X generated by (16), where $\Sigma_{0,ii} = 1$ for all i. First, we define for some $0 < \theta < 1$

$$\mathcal{F}(\theta) := \left\{X : \forall j = 1, \ldots, p, \ 1 - \theta \leq \|X_j\|_2 / \sqrt{n} \leq 1 + \theta\right\}, \tag{34}$$

where $X_1, \ldots, X_p$ are the column vectors of the $n \times p$ design matrix X, which in turn is generated by (16). It is clear when all columns of X have an Euclidean norm close to $\sqrt{n}$, as guaranteed by (34) for some $0 < \theta < 1$ that is small, it makes sense to discuss the RE condition in the form of (35) as formulated in (Bickel et al., 2009). For the integer $1 \leq s_0 \leq s$ as defined in (28) and a positive number k_0, $RE(s_0, k_0, X)$ requires that the following holds for all $\upsilon \neq 0$,

$$\frac{1}{K(s_0, k_0, X)} \stackrel{\triangle}{=} \min_{\substack{J_0 \subset \{1,\ldots,p\}, \\ |J_0| \leq s_0}} \min_{\left\|\upsilon_{J_0^c}\right\|_1 \leq k_0 \left\|\upsilon_{J_0}\right\|_1} \frac{\|X\upsilon\|_2}{\sqrt{n}\left\|\upsilon_{J_0}\right\|_2} > 0, \tag{35}$$

where υ_J represents the subvector of $\upsilon \in \mathbb{R}^p$ confined to a subset J of $\{1, \ldots, p\}$. We now define the following event $\mathcal{R}$ on a random design X generated by (16). which provides an upper bound on $K(s_0, k_0, X)$ for a given $k_0 > 0$ when X satisfies Assumption $RE(s_0, k_0, X)$:

$$\mathcal{R}(\theta) := \left\{X : RE(s_0, k_0, X) \text{ holds with } 0 < K(s_0, k_0, X) \leq \frac{K(s_0, k_0, \Sigma_0)}{1 - \theta}\right\}. \tag{36}$$

Next, for some integer $m < p/2$ to be specified, we define the smallest and largest m-sparse eigenvalues of X generated by (16) to be:

$$\begin{aligned} \Lambda_{\min}(m) &:= \min_{\upsilon \neq 0; m-\text{sparse}} \|X\upsilon\|_2^2 / (n\|\upsilon\|_2^2) \quad \text{and} & (37)\\ \Lambda_{\max}(m) &:= \max_{\upsilon \neq 0; m-\text{sparse}} \|X\upsilon\|_2^2 / (n\|\upsilon\|_2^2). & (38) \end{aligned}$$

Finally, for simplicity, we also define the following event: for $k_0 > 0$ and X as generated by (16),

$$\mathcal{M}(\theta) := \{X : \text{ (40) holds } \forall m \leq \max(s, (k_0 + 1)s_0)\}, \text{ for which} \tag{39}$$

$$0 < (1 - \theta)\sqrt{\rho_{\min}(m)} \leq \sqrt{\Lambda_{\min}(m)} \leq \sqrt{\Lambda_{\max}(m)} \leq (1 + \theta)\sqrt{\rho_{\max}(m)}. \tag{40}$$

Formally, we consider the set of random designs that satisfy all events as defined, for some $0 < \theta < 1$. Theorem 6 shows concentration results that we need for the present work, which follows from Theorem 1.2 and 1.4 in Zhou (2010a).

Theorem 6 *Let $0 < \theta < 1$. Let $\rho_{\min}(s) > 0$, where $s < p/2$ is the maximum node-degree in G. Suppose $RE(s_0, 4, \Sigma_0)$ holds for $s_0 \leq s$ as in (29), where $\Sigma_{0,ii} = 1$ for $i = 1, \ldots, p$. Let $f(m) = \min(4m\rho_{\max}(m)\log(5ep/m), m \log p)$, where $m < p/2$. Let $c', \alpha, \bar{c} > 0$ be some absolute constants and $\bar{C} = (2+k_0)K(s_0, k_0, \Sigma_0)$, where $k_0 > 0$; Suppose the sample size satisfies*

$$n > \frac{9c'\alpha^4}{\theta^2} \max\left(\bar{C}^2 f(s_0), \log p\right) \tag{41}$$

and for $\det_s := \min_T |\Sigma_{0,TT}|$, where $T \subset \{1, \ldots, p\}$ and $|T| = s$

$$n > \frac{18c'\alpha^4}{\theta^2}\left(5s \log 5ep/s + s \log \sqrt{\rho_{\max}(s)} - \frac{1}{2}\log \det_s\right). \tag{42}$$

Then, for a random design X as generated by (16), we have

$$\mathbb{P}(\mathcal{X}) := \mathbb{P}(\mathcal{R}(\theta) \cap \mathcal{F}(\theta) \cap \mathcal{M}(\theta)) \geq 1 - 3\exp(-\bar{c}\theta^2 n/\alpha^4). \tag{43}$$

Clearly we have $\rho_{\max}(s) \leq s$. Thus a sample size of order $n = O(s \log p)$ is sufficient for event $\mathcal{X}$ to hold with probability as in (43), which holds by (A1) as in Section 3 given that $s < p$. We emphasize that we only need the lower bound on n as in (41) if we only aim to obtain (34) and (36); (42) is required to bound sparse eigenvalues of order s as specified in (39).

A.2 Definitions of other various events

Under (A1) as in Section 3, excluding event $\mathcal{X}^c$ as bounded in Theorem 6 and events $\mathcal{C}_a, \mathcal{X}_0$ to be defined in this subsection, we can then proceed to treat $X \in \mathcal{X} \cap \mathcal{C}_a$ as a deterministic design in regression and thresholding, for which $\mathcal{R}(\theta) \cap \mathcal{M}(\theta) \cap \mathcal{F}(\theta)$ holds with $\mathcal{C}_a$, We then make use of event $\mathcal{X}_0$ in the MLE refitting stage for bounding the Frobenius norm. We now define two types of correlations events $\mathcal{C}_a$ and $\mathcal{X}_0$.

Correlation bounds on X_j and V_i. In this section, we first bound the maximum correlation between pairs of random vectors (V_i, X_j), for all i, j where $i \neq j$, each of which corresponds to a pair of variables (V_i, X_j) as defined in (2) and (3). Here we use X_j and V_i, for all i, j, to denote both random vectors and their corresponding variables.

Let us define $\sigma_{V_j} := \sqrt{\mathrm{Var}(V_j)} \geq v > 0$ as a shorthand. Let $V'_j := V_j/\sigma_{V_j}, j = 1, \ldots, p$ be a standard normal random variable. Let us now define for all $j, k \neq j$,

$$Z_{jk} = \frac{1}{n}\langle V'_j, X_k \rangle = \frac{1}{n}\sum_{i=1}^{n} v'_{j,i} x_{k,i},$$

where for all $i = 1, \ldots, n$ $v'_{j,i}, x_{k,i}, \forall j, k \neq j$ are independent standard normal random variables. For some $a \geq 6$, let event

$$\mathcal{C}_a := \left\{\max_{j,k} |Z_{jk}| < \sqrt{1+a}\sqrt{(2\log p)/n} \text{ where } a \geq 6\right\}. \tag{44}$$

Bounds on pairwise correlations in columns of X. Let $\Sigma_0 := (\sigma_{0,ij})$, where we denote $\sigma_{0,ii} := \sigma_i^2$. Denote by $\Delta = X^T X/n - \Sigma_0$. Consider for some constant $C_3 > 4\sqrt{5/3}$,

$$\mathcal{X}_0 := \left\{ \max_{j,k} |\Delta_{jk}| < C_3 \sqrt{\log \max\{p,n\}/n} < 1/2 \right\}. \tag{45}$$

We first state Lemma 7, which is used for bounding a type of correlation events across all regressions; see proof of Theorem 10. It is also clear that event $\mathcal{C}_a$ is equivalent to the event to be defined in (46). Lemma 7 also justifies the choice of λ_n in nodewise regressions (cf. Theorem 10). We then bound event $\mathcal{X}_0$ in Lemma 8. Both proofs appear in Section A.3.

Lemma 7 *Suppose that $p < e^{n/4C_2^2}$. Then with probability at least $1 - 1/p^2$, we have*

$$\forall j \neq k, \quad \left| \frac{1}{n} \langle V_j, X_k \rangle \right| \leq \sigma_{V_j} \sqrt{1+a} \sqrt{(2\log p)/n} \tag{46}$$

where $\sigma_{V_j} = \sqrt{\mathrm{Var}(V_j)}$ and $a \geq 6$. Hence

$$\mathbb{P}(\mathcal{C}_a) \geq 1 - 1/p^2.$$

Lemma 8 *For a random design X as in (1) with $\Sigma_{0,jj} = 1, \forall j \in \{1, \ldots, p\}$, and for $p < e^{n/4C_3^2}$, where $C_3 > 4\sqrt{5/3}$, we have*

$$\mathbb{P}(\mathcal{X}_0) \geq 1 - 1/\max\{n,p\}^2.$$

We note that the upper bounds on p in Lemma 7 and 8 clearly hold given (A1). For the rest of the paper, we prove Theorem 10 in Section B for nodewise regressions. We proceed to derive bounds on selecting an edge set E in Section C. We then derive various bounds on the maximum likelihood estimator given E in Theorem 14- 16 in Section D, where we also prove Theorem 1. Next, we prove Lemma 7 and 8 in Section A.3.

A.3 Proof of Lemma 7 and 8

In this section, we prove Lemma 7 and 8. We first state the following large inequality bound for bounding products of correlated normal random variables.

Lemma 9 Zhou et al. (2008, Lemma 38) *Given a set of identical independent random variables $Y_1, \ldots, Y_n \sim Y$, where $Y = x_1 x_2$, with $x_1, x_2 \sim N(0,1)$ and $\sigma_{12} = \rho_{12}$ with $\rho_{12} \leq 1$ being their correlation coefficient. Let us now define $Q = \frac{1}{n}\sum_{i=1}^n Y_i =: \frac{1}{n}\langle X_1, X_2 \rangle = \frac{1}{n}\sum_{i=1}^n x_{1,i} x_{2,i}$. Let $\Psi_{12} = (1+\sigma_{12}^2)/2$. For $0 \leq \tau \leq \Psi_{12}$,*

$$\mathbb{P}(|Q - \mathbb{E}Q| > \tau) \leq \exp\left\{ -\frac{3n\tau^2}{10(1+\sigma_{12}^2)} \right\} \tag{47}$$

Now by Lemma 9 with $\rho_{jk} = 0, \forall j,k = 1,\ldots,p, j \neq k$ and using the fact that $\mathbb{E}Z_{jk} = 0$, we show Lemma 7.

Proof of Lemma 7. Now it is clear that we have at most $p(p-1)$ unique entries $Z_{jk}, \forall j \neq k$. By the union bound and by taking $\tau = C_2\sqrt{\frac{\log p}{n}}$ in (47) with $\sigma_{jk} = 0, \forall j,k$, we have

$$\begin{aligned} 1 - \mathbb{P}\left(\forall j \neq k, \left|\frac{1}{n}\langle V_j, X_k\rangle\right| \geq \lambda_{\sigma,a,p}\right) &= \mathbb{P}\left(\mathcal{C}_a\right) \\ &\leq \mathbb{P}\left(\max_{jk}|Z_{jk}| \geq C_2\sqrt{\frac{\log p}{n}}\right) \leq (p^2-p)\exp\left(-\frac{3C_2^2\log p}{10}\right) \\ &\leq p^2\exp\left(-\frac{3C_2^2\log p}{10}\right) = p^{-\frac{3C_2^2}{10}+2} < \frac{1}{p^2} \end{aligned}$$

where $\sqrt{2(1+a)} \geq C_2 > 2\sqrt{10/3}$, where $a \geq 6$. Note that $p < e^{n/4C_2^2}$ guarantees that $C_2\sqrt{\frac{\log p}{n}} < 1/2$. ■

In order to bound the probability of event $\mathcal{X}_0$, we first state the following large inequality bound for the non-diagonal entries of Σ_0, which follows immediately from Lemma 9 by plugging in $\sigma_i^2 = \sigma_{0,ii} = 1, \forall i = 1,\ldots,p$ and using the fact that $|\sigma_{0,jk}| = |\rho_{jk}\sigma_j\sigma_k| \leq 1, \forall j \neq k$, where ρ_{jk} is the correlation coefficient between variables X_j and X_k. Let $\Psi_{jk} = (1+\sigma_{0,jk}^2)/2$. Then

$$\mathbb{P}\left(|\Delta_{jk}| > \tau\right) \leq \exp\left\{-\frac{3n\tau^2}{10(1+\sigma_{0,jk}^2)}\right\} \leq \exp\left\{-\frac{3n\tau^2}{20}\right\} \quad \text{for } 0 \leq \tau \leq \Psi_{jk}. \tag{48}$$

We now also state a large deviation bound for the χ_n^2 distribution Johnstone (2001):

$$\mathbb{P}\left(\frac{\chi_n^2}{n} - 1 > \tau\right) \leq \exp\left(\frac{-3n\tau^2}{16}\right), \text{ for } 0 \leq \tau \leq \frac{1}{2}. \tag{49}$$

Lemma 8 follows from (48) and (49) immediately.

Proof of Lemma 8. Now it is clear that we have $p(p-1)/2$ unique non-diagonal entries $\sigma_{0,jk}, \forall j \neq k$ and p diagonal entries. By the union bound and by taking $\tau = C_3\sqrt{\frac{\log\max\{p,n\}}{n}}$ in (49) and (48) with $\sigma_{0,jk} \leq 1$, we have

$$\begin{aligned} \mathbb{P}\left((\mathcal{X}_0)^c\right) &= \mathbb{P}\left(\max_{jk}|\Delta_{jk}| \geq C_3\sqrt{\frac{\log\max\{p,n\}}{n}}\right) \\ &\leq p\exp\left(-\frac{3C_3^2\log\max\{p,n\}}{16}\right) + \frac{p^2-p}{2}\exp\left(-\frac{3C_3^2\log\max\{p,n\}}{20}\right) \\ &\leq p^2\exp\left(-\frac{3C_3^2\log\max\{p,n\}}{20}\right) = (\max\{p,n\})^{-\frac{3C_3^2}{20}+2} < \frac{1}{(\max\{p,n\})^2} \end{aligned}$$

for $C_3 > 4\sqrt{5/3}$, where for the diagonal entries we use (49), and for the non-diagonal entries, we use (48). Finally, $p < e^{n/4C_3^2}$ guarantees that $C_3\sqrt{\frac{\log\max\{p,n\}}{n}} < 1/2$. ■

Appendix B. Bounds for nodewise regressions

In Theorem 10 and Lemma 11 we let s_0^i be as in (28) and T_0^i denote locations of the s_0^i largest coefficients of β^i in absolute values. For the vector h^i to be defined in Theorem 10, we let T_1^i denote the s_0^i largest positions of h^i in absolute values outside of T_0^i; Let $T_{01}^i := T_0^i \cup T_1^i$. We suppress the superscript in T_0^i, T_1^i and thus T_{01}^i throughout this section for clarity.

Theorem 10 (Oracle inequalities of the nodewise regressions) *Let $0 < \theta < 1$. Let $\rho_{\min}(s) > 0$, where $s < p/2$ is the maximum node-degree in G. Suppose $RE(s_0, 4, \Sigma_0)$ holds with $K(s_0, 4, \Sigma_0)$ for $s_0 \leq s$ as defined in (29), where $\Sigma_{0,ii} = 1$ for $i = 1, \ldots, p$. Suppose $\rho_{\max}(\max(s, 3s_0)) < \infty$. The data is generated by $X^{(1)}, \ldots, X^{(n)}$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma_0)$, where n satisfies (41) and (42).*

Consider the nodewise regressions in (10), where for each i, we regress X_i onto the other variables $\{X_k;\ k \neq i\}$ following (2), where $V_i \sim N(0, \mathrm{Var}(V_i))$ is independent of $X_j, \forall j \neq i$ as in (3) and

$$\mathrm{Var}(V_i) = 1/\theta_{0,ii} \ \textit{ iff } \beta_j^i = -\theta_{0,ij}/\theta_{0,ii}.$$

Let β_{init}^i be an optimal solution to (10) for each i. Let $\lambda_n = d_0 \lambda = d_0^i \lambda \sigma_{V_i}$ where d_0 is chosen such that $d_0 \geq 2(1+\theta)\sqrt{1+a}$ holds for some $a \geq 6$, and clearly $d_0^i \geq d_0$. Let $h^i = \beta_{\text{init}}^i - \beta_{T_0}^i$. Then simultaneously for all i, on $\mathcal{C}_a \cap \mathcal{X}$, where $\mathcal{X} := \mathcal{R}(\theta) \cap \mathcal{F}(\theta) \cap \mathcal{M}(\theta)$, we have

$$\begin{aligned}
\left\|\beta_{\text{init}}^i - \beta^i\right\|_2 &\leq \lambda \sqrt{s_0^i} d_0 \sqrt{2D_0^2 + 2D_1^2 + 2}, \ \textit{where} \\
\|h_{T_{01}}\|_2 &\leq D_0 d_0 \lambda \sqrt{s_0^i} \ \textit{and} \qquad (50)\\
\left\|h_{T_0^c}^i\right\|_1 = \left\|\beta_{\text{init},T_0^c}^i\right\|_1 &\leq D_1 d_0 \lambda s_0^i \qquad (51)
\end{aligned}$$

where D_0, D_1 are defined in (88) and (90) respectively.

The choice of d_0 will be justified in Section E, where we also calculate D_0, D_1 to be shown as in (52). Suppose we choose for some constant $c_0 \geq 4\sqrt{2}$ and $a_0 = 7$,

$$d_0 = c_0(1+\theta)^2 \sqrt{\rho_{\max}(s)\rho_{\max}(3s_0)},$$

where we assume that $\rho_{\max}(\max(s, 3s_0)) < \infty$ are reasonably bounded,then

$$D_0 \leq \frac{5K^2(s_0, 4, \Sigma_0)}{(1-\theta)^2} \text{ and } D_1 \leq \frac{49K^2(s_0, 4, \Sigma_0)}{16(1-\theta)^2}. \qquad (52)$$

Proof Consider each regression function in (10) with $X_{\cdot \setminus i}$ being the design matrix and X_i the response vector, where $X_{\cdot \setminus i}$ denotes columns of X excluding X_i. It is clear that for $\lambda_n = d_0 \lambda$, we have for $i = 1, \ldots, p$ and $a \geq 6$,

$$\lambda_n = (d_0/\sigma_{V_i})\sigma_{V_i}\lambda := d_0^i \sigma_{V_i}\lambda \geq d_0 \lambda \sigma_{V_i} \geq 2(1+\theta)\lambda\sqrt{1+a}\sigma_{V_i} = 2(1+\theta)\lambda_{\sigma,a,p}$$

such that (87) holds given that $\sigma_{V_i} \leq 1, \forall i$, where it is understood that $\sigma := \sigma_{V_i}$.

It is also clear that on $\mathcal{C}_a \cap \mathcal{X}$, event $\mathcal{T}_a \cap \mathcal{X}$ holds for each regression when we invoke Theorem 25, with $Y := X_i$ and $X := X_{\cdot \setminus i}$, for $i = 1, \ldots, p$. We can then invoke bounds for each individual regression as in Theorem 25, and conclude that the present theorem holds, noting that $d_0^i \sigma_{V_i} = d_0$ by definition. ■

Appendix C. Bounds on thresholding

In this section, we first show Lemma 11, following conditions in Theorem 10. We then show Corollary 12, which proves Proposition 2 and the first statement of Theorem 1.

Lemma 11 *Suppose $RE(s_0, 4, \Sigma_0)$ holds for s_0 be as in (29) and $\rho_{\min}(s) > 0$, where $s < p/2$ is the maximum node-degree in G. Suppose $\rho_{\max}(\max(s, 3s_0)) < \infty$. Let $S^i = \{j : j \neq i : |\beta_j^i \neq 0\}$. Let $c_0 \geq 4\sqrt{2}$ be some absolute constant. Suppose n satisfies (41) and (42). Let β_{init}^i be an optimal solution to (10) with*

$$\lambda_n = d_0 \lambda \;\; \text{where} \;\; d_0 = c_0 (1+\theta)^2 \sqrt{\rho_{\max}(s)\rho_{\max}(3s_0)};$$

Suppose for each regression, we apply the same threshold rule to obtain a subset I^i as follows,

$$I^i = \{j : j \neq i, \; \left|\beta_{j,\text{init}}^i\right| \geq t_0 = f_0 \lambda\}, \;\; \text{and} \;\; \mathcal{D}^i := \{1, \ldots, i-1, i+1, \ldots, p\} \setminus I^i$$

where $f_0 := D_4 d_0$ for some constant D_4 to be specified. Then we have on event $\mathcal{C}_a \cap \mathcal{X}$,

$$\begin{aligned} |I^i| &\leq s_0^i (1 + D_1/D_4) \text{ and } \;\; |I^i \cup S^i| \leq s^i + (D_1/D_4) s_0^i \;\; \text{and} & (53)\\ \left\|\beta_{\mathcal{D}}^i\right\|_2 &\leq d_0 \lambda \sqrt{s_0^i} \sqrt{1 + (D_0 + D_4)^2} & (54)\end{aligned}$$

where $\mathcal{D}$ is understood to be $\mathcal{D}^i$ and D_0, D_1 are understood to be the same constants as in Theorem 10.

Proof Let $T_0 := T_0^i$ denote the s_0^i largest coefficients of β^i in absolute values. By (51), we have for $f_0 = D_4 d_0$

$$|I^i \cap T_0^c| \leq \left\|\beta_{\text{init}, T_0^c}^i\right\|_1 \frac{1}{f_0 \lambda} \leq D_1 d_0 s_0^i / (D_4 d_0) \leq D_1 s_0^i / D_4 \tag{55}$$

where D_1 is understood to be the same constant that appears in (51). Thus we have

$$\left|I^i\right| = |I^i \cap T_0^c| + |I^i \cap T_0| \leq s_0^i (1 + D_1/D_4).$$

Now the second inequality in (53) clearly holds given (55) and the following:

$$|I^i \cup S^i| \leq |S^i| + |I^i \cap (S^i)^c| \leq s^i + |I^i \cap (T_0^i)^c|.$$

We now bound $\left\|\beta_{\mathcal{D}}^i\right\|_2^2$ following essentially the arguments as in Zhou (2010b). We have

$$\left\|\beta_{\mathcal{D}}^i\right\|_2^2 = \left\|\beta_{T_0 \cap \mathcal{D}}^i\right\|_2^2 + \left\|\beta_{T_0^c \cap \mathcal{D}}^i\right\|_2^2,$$

where the second term is bounded as: $\left\|\beta^i_{T_0^c \cap \mathcal{D}}\right\|_2^2 \le \left\|\beta^i_{T_0^c}\right\|_2^2 \le s_0^i \lambda^2 \sigma^2_{V_i}$ by definition of s_0^i as in (28) and (32); For the first term, we have by the triangle inequality,

$$\begin{aligned}
\left\|\beta^i_{T_0 \cap \mathcal{D}}\right\|_2 &\le \left\|(\beta^i - \beta^i_{\text{init}})_{T_0 \cap \mathcal{D}}\right\|_2 + \left\|(\beta^i_{\text{init}})_{T_0 \cap \mathcal{D}}\right\|_2 \\
&\le \left\|(\beta^i - \beta^i_{\text{init}})_{T_0}\right\|_2 + t_0\sqrt{|T_0 \cap \mathcal{D}|} \le \|h_{T_0}\|_2 + t_0\sqrt{s_0^i} \\
&\le D_0 d_0 \lambda \sqrt{s_0^i} + D_4 d_0 \lambda \sqrt{s_0^i} \le (D_0 + D_4) d_0 \lambda \sqrt{s_0^i}
\end{aligned}$$

where we invoked the bound on $\|h_{T_0}\|_2$ as in (50) following Theorem 10. Thus we have (54). ■

Recall $\Theta_0 = \Sigma_0^{-1}$. Let $\Theta_{0,\mathcal{D}}$ denote the submatrix of Θ_0 indexed by $\mathcal{D}$ as in (21) with all other positions set to be 0. Let E_0 be the true edge set.

Corollary 12 *Suppose all conditions in Lemma 11 hold. Then on event $\mathcal{C}_a \cap \mathcal{X}$, for $\widetilde{\Theta}_0$ as in (23) and E as in (22), we have for $S_{0,n}$ as in (30) and $\Theta_0 = (\theta_{0,ij})$*

$$|E| \le 2(1 + D_1/D_4)S_{0,n} \;\; \textit{where} \;\; |E \setminus E_0| \le 2D_1/D_4 S_{0,n} \tag{56}$$

$$\begin{aligned}
\|\Theta_{0,\mathcal{D}}\|_F &:= \left\|\widetilde{\Theta}_0 - \Theta_0\right\|_F \\
&\le \sqrt{\min\{S_{0,n}(\max_{i=1,\ldots p} \theta^2_{0,ii}), s_0 \|\text{diag}(\Theta_0)\|_F^2\}} \sqrt{(1 + (D_0 + D_4)^2)} d_0 \lambda \qquad (57) \\
&:= \sqrt{S_{0,n}\,(1 + (D_0 + D_4)^2) C_{\text{diag}}} d_0 \lambda
\end{aligned}$$

where $C^2_{\text{diag}} := \min\{\max_{i=1,\ldots p} \theta^2_{0,ii}, (s_0/S_{0,n}) \|\text{diag}(\Theta_0)\|_F^2\}$, and D_0, D_1 are understood to be the same constants as in Theorem 10. Clearly, for $D_4 \ge D_1$, we have (19).

Proof It is clear that by the OR rule in (9), which will allow either $\widehat{\beta}^i_j$ or $\widehat{\beta}^j_i$ to be non-zero to turn it on; and hence we could turn on at most $2|I_i|$ edges. These rule will allow us to keep more edges that we would have reduced by the thresholding rule at each node. We have by (53)

$$|E| \le \sum_{i=1,\ldots p} 2(1 + D_1/D_4)s_0^i = 2\,(1 + D_1/D_4)\,S_{0,n},$$

where $(2D_1/D_4)S_{0,n}$ is an upper bound on $|E \setminus E_0|$ by (55). We now obtain a bound on $\|\Theta_{0,\mathcal{D}}\|_F^2$ as follows:

$$\begin{aligned}
\|\Theta_{0,\mathcal{D}}\|_F^2 &\le \sum_{i=1}^{p} \theta^2_{0,ii} \left\|\beta^i_{\mathcal{D}}\right\|_2^2 \le (1 + (D_0 + D_4)^2) d_0^2 \lambda^2 \sum_{i=1}^{p} \theta^2_{0,ii} s_0^i \\
&\le \min\{S_{0,n}(\max_{i=1,\ldots p} \theta^2_{0,ii}), s_0 \|\text{diag}(\Theta_0)\|_F^2\}(1 + (D_0 + D_4)^2) d_0^2 \lambda^2
\end{aligned}$$

■

Remark 13 *Recall we assume that* $\max\{p, S_{0,n}\} = o(n/\log(p))$ *in Proposition 2, which guarantees that* $n = \Omega(p \log p)$ *holds as required by Theorem 10 and Lemma 11 in the worst case scenario. Note that if* s_0 *is small, then the second term in* C_{diag} *will provide a tighter bound.*

Appendix D. Bounds on MLE refitting

To facilitate technical discussions, we need to introduce some more notation. Let $\mathcal{S}^p_{++}$ denote the set of $p \times p$ symmetric positive definite matrices:

$$\mathcal{S}^p_{++} = \{\Theta \in \mathbb{R}^{p\times p} | \Theta \succ 0\}.$$

Let us define a subspace $\mathcal{S}^p_E$ corresponding to an edge set $E \subset \{(i,j) : i,j = 1,\ldots,p, i \neq j\}$:

$$\begin{aligned} \mathcal{S}^p_E &:= \{\Theta \in \mathbb{R}^{p\times p} | \theta_{0,ij} = 0 \text{ for all } i \neq j \text{ such that } (i,j) \notin E\} \\ \text{and } \mathcal{S}_n &= \{\Theta : \Theta \in \mathcal{S}^p_{++} \cap \mathcal{S}^p_E\}. \end{aligned} \tag{58}$$

Recall the maximum likelihood estimate $\widehat{\Theta}_n$ as in (59) minimizes over all $\Theta \in \mathcal{S}_n$ the empirical risk as in expression (24). Thus, we write

$$\widehat{\Theta}_n(E) := \arg\min_{\Theta\in\mathcal{S}_n} \widehat{R}_n(\Theta) = \arg \min_{\Theta \in \mathcal{S}^p_{++}\cap\mathcal{S}^p_E} \left\{\mathrm{tr}(\Theta\widehat{S}_n) - \log|\Theta|\right\} \tag{59}$$

which gives the "best" refitted sparse estimator given a sparse subset of edges E that we obtain from the nodewise regressions and thresholding. We note that the estimator (59) remains to be a convex optimization problem, as the constraint set is the intersection the positive definite cone $\mathcal{S}^p_{++}$ and the linear subspace $\mathcal{S}^p_E$. It is not hard to see that the estimator (59) is equivalent to (13).

Theorem 14 *Consider data generating random variables as in expression (16) and assume that* $(A1)$, (26), *and* (27) *hold. Let* $\mathcal{E}$ *be some event such that* $\mathbb{P}(\mathcal{E}) \geq 1 - d/p^2$ *for a small constant* d. *Suppose on event* $\mathcal{E}$*:*

1. *We obtain an edge set* E *such that its size* $|E| = lin(S_{0,n}, p)$ *is a linear function in* $S_{0,n}$ *and* p*, where* $S_{0,n}$ *is as defined in* (30)*;*
2. *And for* $\widetilde{\Theta}_0$ *as in* (23) *and for some constant* C_{bias} *to be specified, we have*

$$\|\Theta_{0,\mathcal{D}}\|_F := \left\|\widetilde{\Theta}_0 - \Theta_0\right\|_F \leq C_{\text{bias}}\sqrt{2S_{0,n}\log(p)/n} < \underline{c}/32. \tag{60}$$

Let $\widehat{\Theta}_n(E)$ *be as defined in* (59)*. Suppose the sample size satisfies for* $C_3 \geq 4\sqrt{5/3}$*,*

$$n > \frac{106}{\underline{k}^2}\left(C_3 + \frac{32}{31\underline{c}^2}\right)^2 \max\left\{(p + 2|E|)\log(n),\ C^2_{\text{bias}} 2S_{0,n}\log p\right\}. \tag{61}$$

Then on event $\mathcal{E} \cap \mathcal{X}_0$*, we have for* $M = (9/(2\underline{k}^2)) \cdot \left(C_3 + 32/(31\underline{c}^2)\right)$

$$\left\|\widehat{\Theta}_n(E) - \Theta_0\right\|_F \leq (M+1)\max\left\{\sqrt{(p+2|E|)\log(n)/n},\ C_{\text{bias}}\sqrt{2S_{0,n}\log(p)/n}\right\}. \tag{62}$$

We note that although Theorem 14 is meant for proving Theorem 1, we state it as an independent result; For example, one can indeed take E from Corollary 12, where we have $|E| \leq cS_{0,n}$ for some constant c for $D_4 \asymp D_1$. In view of (57), we aim to recover $\widetilde{\Theta}_0$ by $\widehat{\Theta}_n(E)$ as defined in (59). In Section D.2, we will focus in Theorem 14 on bounding for W suitably chosen,

$$\left\|\widehat{\Theta}_n(E) - \widetilde{\Theta}_0\right\|_F = O_P\left(W\sqrt{(p + S_{0,n})\log(n)/n}\right).$$

By the triangle inequality, we conclude that

$$\left\|\widehat{\Theta}_n(E) - \Theta_0\right\|_F \leq \left\|\widehat{\Theta}_n(E) - \widetilde{\Theta}_0\right\|_F + \left\|\widetilde{\Theta}_0 - \Theta_0\right\|_F = O_P\left(W\sqrt{(p + S_{0,n})\log(n)/n}\right).$$

We now state bounds for the convergence rate on Frobenius norm of the covariance matrix and for KL divergence. We note that constants have not been optimized. Proofs of Theorem 15 and 16 appear in Section D.3 and D.4 respectively.

Theorem 15 *Suppose all conditions, events, and bounds on $|E|$ and $\|\Theta_{0,\mathcal{D}}\|_F$ in Theorem 14 hold. Let $\widehat{\Theta}_n(E)$ be as defined in (59). Suppose the sample size satisfies for $C_3 \geq 4\sqrt{5/3}$ and C_{bias}, M as defined in Theorem 14*

$$n > \frac{106}{\underline{c}^2\underline{k}^4}\left(C_3 + \frac{32}{31\underline{c}^2}\right)^2 \max\left\{(p + 2|E|)\log(n),\ C_{\text{bias}}^2 2S_{0,n}\log p\right\}. \tag{63}$$

Then on event $\mathcal{E} \cap \mathcal{X}_0$, we have $\varphi_{\min}(\widehat{\Theta}_n(E)) > \underline{c}/2 > 0$ and for $\widehat{\Sigma}_n(E) = (\widehat{\Theta}_n(E))^{-1}$,

$$\left\|\widehat{\Sigma}_n(E) - \Sigma_0\right\|_F \leq \frac{2(M+1)}{\underline{c}^2}\max\left\{\sqrt{\frac{(p + 2|E|)\log(n)}{n}},\ C_{\text{bias}}\sqrt{\frac{2S_{0,n}\log(p)}{n}}\right\}. \tag{64}$$

Theorem 16 *Suppose all conditions, events, and bounds on $|E|$ and $\|\Theta_{0,\mathcal{D}}\|_F := \left\|\widetilde{\Theta}_0 - \Theta_0\right\|_F$ in Theorem 14 hold. Let $\widehat{\Theta}_n(E)$ be as defined in (59). Suppose the sample size satisfies (61) for $C_3 \geq 4\sqrt{5/3}$ and C_{bias}, M as defined in Theorem 14. Then on event $\mathcal{E} \cap \mathcal{X}_0$, we have for $R(\widehat{\Theta}_n(E)) - R(\Theta_0) \geq 0$,*

$$R(\widehat{\Theta}_n(E)) - R(\Theta_0) \leq M(C_3 + 1/8)\max\left\{(p + 2|E|)\log(n)/n,\ C_{\text{bias}}^2 2S_{0,n}\log(p)/n\right\}. \tag{65}$$

D.1 Proof of Theorem 1

Clearly the sample requirements as in (41), (42) are satisfied for some $\theta > 0$ that is appropriately chosen, given (61). In view of Corollary 12, we have on $\mathcal{E} := \mathcal{X} \cap \mathcal{C}_a$: for C_{diag} as in (18)

$$\begin{aligned}
|E| &\leq 2(1 + \frac{D_1}{D_4})S_{0,n} \leq 4S_{0,n} \text{ for } D_4 \geq D_1 \text{ and} \\
\|\Theta_{0,\mathcal{D}}\|_F &:= \left\|\widetilde{\Theta}_0 - \Theta_0\right\|_F \leq C_{\text{bias}}\sqrt{2S_{0,n}\log(p)/n} \leq \underline{c}/32 \text{ where} \\
C_{\text{bias}}^2 &:= \min\left\{\max_{i=1,\ldots p}\theta_{0,ii}^2,\ \frac{s_0}{S_{0,n}}\|\text{diag}(\Theta_0)\|_F^2\right\} d_0^2(1 + (D_0 + D_4)^2) \\
&= C_{\text{diag}}^2 d_0^2(1 + (D_0 + D_4)^2)
\end{aligned} \tag{66}$$

Clearly the last inequality in (60) hold so long as $n > 32^2 C_{\text{bias}}^2 2S_{0,n}\log(p)/\underline{c}^2$, which holds given (61). Plugging in $|E|$ in (62), we have on $\mathcal{E}\cap\mathcal{X}_0$,

$$\left\|\widehat{\Theta}_n(E)-\Theta_0\right\|_F \le (M+1)\max\left\{\sqrt{\frac{(p+4(1+D_1/D_4)S_{0,n})\log(n)}{n}},\ C_{\text{bias}}\sqrt{\frac{2S_{0,n}\log p}{n}}\right\}$$

Now if we take $D_4 \ge D_1$, then we have (19) on event $\mathcal{E}$; and moreover on $\mathcal{E}\cap\mathcal{X}_0$,

$$\begin{aligned}\left\|\widehat{\Theta}_n(E)-\Theta_0\right\|_F &\le (M+1)\max\left\{\sqrt{(p+8S_{0,n})\log(n)/n},\ C_{\text{bias}}\sqrt{2S_{0,n}\log(p)/n}\right\}\\ &\le W\sqrt{(p+S_{0,n})\log(n)/n}\end{aligned}$$

where $W \le \sqrt{2}(M+1)\max\{C_{\text{diag}}d_0\sqrt{1+(D_0+D_4)^2},2\}$. Similarly, we get the bound on $\left\|\widehat{\Sigma}_n-\Sigma_0\right\|_F$ with Theorem 15, and the bound on risk following Theorem 16. Thus all statements in Theorem 1 hold. ■

Remark 17 *Suppose event $\mathcal{E}\cap\mathcal{X}_0$ holds. Now suppose that we take $D_4=1$, that is, if we take the threshold to be exactly the penalty parameter λ_n:*

$$t_0 = d_0\lambda := \lambda_n.$$

Then we have on event $\mathcal{E}$ by (56) $|E| \le 2(1+D_1)S_{0,n}$ and $|E\setminus E_0| \le 2D_1S_{0,n}$ and on event on $\mathcal{E}\cap\mathcal{X}_0$, for $C'_{\text{bias}} := C_{\text{diag}}d_0\sqrt{1+(D_0+1)^2}$

$$\left\|\widehat{\Theta}_n(E)-\Theta_0\right\|_F \le M\max\left\{\sqrt{\frac{(p+4(1+D_1)S_{0,n})\log(n)}{n}},\ C'_{\text{bias}}\sqrt{\frac{2S_{0,n}\log p}{n}}\right\}$$

It is not hard to see that we achieve essential the same rate as stated in Theorem 1, with perhaps slightly more edges included in E.

D.2 Proof of Theorem 14

Suppose event $\mathcal{E}$ holds throughout this proof. We first obtain the bound on spectrum of $\widetilde{\Theta}_0$: It is clear that by (26) and (60), we have on $\mathcal{E}$,

$$\varphi_{\min}(\widetilde{\Theta}_0) \ \ge\ \varphi_{\min}(\Theta_0)-\left\|\widetilde{\Theta}_0-\Theta_0\right\|_2 \ge \varphi_{\min}(\Theta_0)-\|\Theta_{0,\mathcal{D}}\|_F > 31\underline{c}/32, \tag{67}$$

$$\varphi_{\max}(\widetilde{\Theta}_0) \ <\ \varphi_{\max}(\Theta_0)+\left\|\widetilde{\Theta}_0-\Theta_0\right\|_2 \le \varphi_{\max}(\Theta_0)+\|\Theta_{0,\mathcal{D}}\|_F < \frac{\underline{c}}{32}+\frac{1}{\underline{k}}. \tag{68}$$

Throughout this proof, we let $\Sigma_0 = (\sigma_{0,ij}) := \Theta_0^{-1}$. In view of (67), define $\widetilde{\Sigma}_0 := (\widetilde{\Theta}_0)^{-1}$. We use $\widehat{\Theta}_n := \widehat{\Theta}_n(E)$ as a shorthand.

Given $\widetilde{\Theta}_0 \in \mathcal{S}^p_{++}\cap\mathcal{S}^p_E$ as guaranteed in (67), let us define a new convex set:

$$U_n(\widetilde{\Theta}_0) := (\mathcal{S}^p_{++}\cap\mathcal{S}^p_E)-\widetilde{\Theta}_0 = \{B-\widetilde{\Theta}_0 | B\in\mathcal{S}^p_{++}\cap\mathcal{S}^p_E\}\subset\mathcal{S}^p_E$$

which is a translation of the original convex set $\mathcal{S}^p_{++} \cap \mathcal{S}^p_E$. Let $\underline{0}$ be a matrix with all entries being zero. Thus it is clear that $U_n(\widetilde{\Theta}_0) \ni \underline{0}$ given that $\widetilde{\Theta}_0 \in \mathcal{S}^p_{++} \cap \mathcal{S}^p_E$. Define for $\widehat{R}_n$ as in expression (24),

$$\begin{aligned}\widetilde{Q}(\Theta) &:= \widehat{R}_n(\Theta) - \widehat{R}_n(\widetilde{\Theta}_0) = \mathrm{tr}(\Theta \widehat{S}_n) - \log|\Theta| - \mathrm{tr}(\widetilde{\Theta}_0 \widehat{S}_n) + \log|\widetilde{\Theta}_0| \\ &= \mathrm{tr}\left((\Theta - \widetilde{\Theta}_0)(\widehat{S}_n - \widetilde{\Sigma}_0)\right) - (\log|\Theta| - \log|\widetilde{\Theta}_0|) + \mathrm{tr}\left((\Theta - \widetilde{\Theta}_0)\widetilde{\Sigma}_0\right).\end{aligned}$$

For an appropriately chosen r_n and a large enough $M > 0$, let

$$\mathbb{T}_n = \{\Delta \in U_n(\widetilde{\Theta}_0), \|\Delta\|_F = M r_n\}, \text{ and} \tag{69}$$

$$\Pi_n = \{\Delta \in U_n(\widetilde{\Theta}_0), \|\Delta\|_F < M r_n\}. \tag{70}$$

It is clear that both Π_n and $\mathbb{T}_n \cup \Pi_n$ are convex. It is also clear that $\underline{0} \in \Pi_n$. Throughout this section, we let

$$r_n = \max\left\{\sqrt{\frac{(p + 2|E|)\log(n)}{n}}, C_{\text{bias}}\sqrt{\frac{2S_{0,n}\log p}{n}}\right\}. \tag{71}$$

Define for $\Delta \in U_n(\widetilde{\Theta}_0)$,

$$\widetilde{G}(\Delta) := \widetilde{Q}(\widetilde{\Theta}_0 + \Delta) = \mathrm{tr}(\Delta(\widehat{S}_n - \widetilde{\Sigma}_0)) - (\log|\widetilde{\Theta}_0 + \Delta| - \log|\widetilde{\Theta}_0|) + \mathrm{tr}(\Delta\widetilde{\Sigma}_0) \tag{72}$$

It is clear that $\widetilde{G}(\Delta)$ is a convex function on $U_n(\widetilde{\Theta}_0)$ and $\widetilde{G}(\underline{0}) = \widetilde{Q}(\widetilde{\Theta}_0) = 0$.

Now, $\widehat{\Theta}_n$ minimizes $\widetilde{Q}(\Theta)$, or equivalently $\widehat{\Delta} = \widehat{\Theta}_n - \widetilde{\Theta}_0$ minimizes $\widetilde{G}(\Delta)$. Hence by definition,

$$\widetilde{G}(\widehat{\Delta}) \leq \widetilde{G}(\underline{0}) = 0$$

Note that $\mathbb{T}_n$ is non-empty, while clearly $\underline{0} \in \Pi_n$. Indeed, consider $B_\epsilon := (1+\epsilon)\widetilde{\Theta}_0$, where $\epsilon > 0$; it is clear that $B_\epsilon - \widetilde{\Theta}_0 \in \mathcal{S}^p_{++} \cap \mathcal{S}^p_E$ and $\left\|B_\epsilon - \widetilde{\Theta}_0\right\|_F = |\epsilon| \left\|\widetilde{\Theta}_0\right\|_F = M r_n$ for $|\epsilon| = M r_n / \left\|\widetilde{\Theta}_0\right\|_F$. Note also if $\Delta \in \mathbb{T}_n$, then $\Delta_{ij} = 0 \forall (i, j : i \neq j) \notin E$; Thus we have $\Delta \in \mathcal{S}^p_E$ and

$$\|\Delta\|_0 \leq p + 2|E| \leq p + 2\,\mathrm{lin}(S_{0,n}, p). \tag{73}$$

We now show the following two propositions. Proposition 18 follows from standard results.

Proposition 18 *Let B be a $p \times p$ matrix. If $B \succ 0$ and $B + D \succ 0$, then $B + vD \succ 0$ for all $v \in [0, 1]$.*

Proposition 19 *Under (26), we have for all $\Delta \in \mathbb{T}_n$ such that $\|\Delta\|_F = M r_n$ for r_n as in (71), $\widetilde{\Theta}_0 + v\Delta \succ 0, \forall v \in$ an open interval $I \supset [0, 1]$ on event $\mathcal{E}$.*

Proof In view of Proposition 18, it is sufficient to show that $\widetilde{\Theta}_0 + (1 + \varepsilon)\Delta, \widetilde{\Theta}_0 - \varepsilon\Delta \succ 0$ for some $\varepsilon > 0$. Indeed, by definition of $\Delta \in \mathbb{T}_n$, we have $\varphi_{\min}(\widetilde{\Theta}_0 + \Delta) \succ 0$ on event $\mathcal{E}$; thus

$$\begin{aligned}\varphi_{\min}(\widetilde{\Theta}_0 + (1 + \varepsilon)\Delta) &\geq \varphi_{\min}(\widetilde{\Theta}_0 + \Delta) - \varepsilon\|\Delta\|_2 > 0 \\ \text{and } \varphi_{\min}(\widetilde{\Theta}_0 - \varepsilon\Delta) &\geq \varphi_{\min}(\widetilde{\Theta}_0) - \varepsilon\|\Delta\|_2 > 31\underline{c}/32 - \varepsilon\|\Delta\|_2 > 0\end{aligned}$$

for $\varepsilon > 0$ that is sufficiently small. ■

Thus we have that $\log|\widetilde{\Theta}_0 + v\Delta|$ is infinitely differentiable on the open interval $I \supset [0,1]$ of v. This allows us to use the Taylor's formula with integral remainder to obtain the following:

Lemma 20 *On event $\mathcal{E} \cap \mathcal{X}_0$, $\widetilde{G}(\Delta) > 0$ for all $\Delta \in \mathbb{T}_n$.*

Proof Let us use $\widetilde{A}$ as a shorthand for

$$\mathrm{vec}\Delta^T \left(\int_0^1 (1-v)(\widetilde{\Theta}_0 + v\Delta)^{-1} \otimes (\widetilde{\Theta}_0 + v\Delta)^{-1} dv \right) \mathrm{vec}\Delta,$$

where $\otimes$ is the Kronecker product (if $W = (w_{ij})_{m\times n}$, $P = (b_{k\ell})_{p\times q}$, then $W \otimes P = (w_{ij}P)_{mp\times nq}$), and $\mathrm{vec}\Delta \in \mathbb{R}^{p^2}$ is $\Delta_{p\times p}$ vectorized. Now, the Taylor expansion gives for all $\Delta \in \mathbb{T}_n$,

$$\begin{aligned} \log|\widetilde{\Theta}_0 + \Delta| - \log|\widetilde{\Theta}_0| &= \frac{d}{dv}\log|\widetilde{\Theta}_0 + v\Delta||_{v=0}\Delta + \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\widetilde{\Theta}_0 + v\Delta|dv \\ &= \mathrm{tr}(\widetilde{\Sigma}_0\Delta) - \widetilde{A}, \end{aligned}$$

where $\mathrm{tr}(\widetilde{\Sigma}_0\Delta) = \mathrm{tr}((\Theta - \widetilde{\Theta}_0)\widetilde{\Sigma}_0)$. Hence for all $\Delta \in \mathbb{T}_n$,

$$\widetilde{G}(\Delta) = \widetilde{A} + \mathrm{tr}\left(\Delta(\widehat{S}_n - \widetilde{\Sigma}_0)\right) = \widetilde{A} + \mathrm{tr}\left(\Delta(\widehat{S}_n - \Sigma_0)\right) - \mathrm{tr}\left(\Delta(\widetilde{\Sigma}_0 - \Sigma_0)\right) \tag{74}$$

where we first bound $\mathrm{tr}(\Delta(\widetilde{\Sigma}_0 - \Sigma_0))$ as follows: by (60) and (67), we have on event $\mathcal{E}$

$$\begin{aligned} \left|\mathrm{tr}(\Delta(\widetilde{\Sigma}_0 - \Sigma_0))\right| &= \left|\langle\, \Delta, (\widetilde{\Sigma}_0 - \Sigma_0)\,\rangle\right| \le \|\Delta\|_F \left\|\widetilde{\Sigma}_0 - \Sigma_0\right\|_F \\ &\le \|\Delta\|_F \frac{\|\Theta_{0,\mathcal{D}}\|_F}{\varphi_{\min}(\widetilde{\Theta}_0)\varphi_{\min}(\Theta_0)} \\ &< \|\Delta\|_F \frac{32C_{\text{bias}}\sqrt{2S_{0,n}\log p/n}}{31\underline{c}^2} \le \|\Delta\|_F \frac{32r_n}{31\underline{c}^2}. \end{aligned} \tag{75}$$

Now, conditioned on event $\mathcal{X}_0$, by Lemma 8 and (61)

$$\max_{j,k} |\widehat{S}_{n,jk} - \sigma_{0,jk}| \le C_3\sqrt{\log(n)/n} =: \delta_n$$

and thus with probability $1 - \frac{1}{n^2}$ we have $\left|\mathrm{tr}\left(\Delta(\widehat{S}_n - \Sigma_0)\right)\right| \le \delta_n\,|\Delta|_1$; hence by Cauchy-Schwartz and (73), we have on event $\mathcal{E} \cap \mathcal{X}_0$,

$$\begin{aligned} \mathrm{tr}\left(\Delta(\widehat{S}_n - \Sigma_0)\right) &\ge -\delta_n\,|\Delta|_1 \ge -\delta_n\sqrt{\|\Delta\|_0}\,\|\Delta\|_F \\ &\ge -\delta_n\sqrt{p + 2|E|}\,\|\Delta\|_F \ge -C_3 r_n\,\|\Delta\|_F\,. \end{aligned} \tag{76}$$

Finally, we bound $\widetilde{A}$. First we note that for $\Delta \in \mathbb{T}_n$, we have on event $\mathcal{E}$,

$$\|\Delta\|_2 \le \|\Delta\|_F = Mr_n < \frac{7}{16\underline{k}}, \tag{77}$$

given (61): $n > (\frac{16}{7}\cdot\frac{9}{2\underline{k}})^2 \left(C_3 + \frac{32}{31\underline{c}^2}\right)^2 \max\left\{(p+2|E|)\log(n),\ C_{\text{bias}}^2 2S_{0,n}\log p\right\}$. Now we have by (68) and (27) following Rothman et al. (2008) (see Page 502, proof of Theorem 1 therein): on event $\mathcal{E}$,

$$\begin{aligned}
\widetilde{A} &\geq \|\Delta\|_F^2 \,/\, \left(2\left(\varphi_{\max}(\widetilde{\Theta}_0) + \|\Delta\|_2\right)^2\right) \\
&\geq \|\Delta\|_F^2 \,/\, \left(2(\frac{1}{\underline{k}} + \frac{\underline{c}}{32} + \frac{7}{16\underline{k}})^2\right) > \|\Delta\|_F^2 \frac{2\underline{k}^2}{9} \qquad (78)
\end{aligned}$$

Now on event $\mathcal{E}\cap\mathcal{X}_0$, for all $\Delta \in \mathbb{T}_n$, we have by (74),(78), (76), and (75),

$$\begin{aligned}
\widetilde{G}(\Delta) &> \|\Delta\|_F^2 \frac{2\underline{k}^2}{9} - C_3 r_n \|\Delta\|_F - \|\Delta\|_F \frac{32 r_n}{31\underline{c}^2} \\
&= \|\Delta\|_F^2 \left(\frac{2\underline{k}^2}{9} - \frac{1}{\|\Delta\|_F}\left(C_3 r_n + \frac{32 r_n}{31\underline{c}^2}\right)\right) \\
&= \|\Delta\|_F^2 \left(\frac{2\underline{k}^2}{9} - \frac{1}{M}\left(C_3 + \frac{32}{31\underline{c}^2}\right)\right)
\end{aligned}$$

hence we have $\widetilde{G}(\Delta) > 0$ for M large enough, in particular $M = (9/(2\underline{k}^2))\left(C_3 + 32/(31\underline{c}^2)\right)$ suffices. ■

We next state Proposition 21, which follows exactly that of Claim 12 of Zhou et al. (2008).

Proposition 21 *Suppose event $\mathcal{E}$ holds. If $\widetilde{G}(\Delta) > 0, \forall \Delta \in \mathbb{T}_n$, then $\widetilde{G}(\Delta) > 0$ for all Δ in*

$$\mathbb{W}_n = \{\Delta : \Delta \in U_n(\widetilde{\Theta}_0), \|\Delta\|_F > M r_n\}$$

for r_n as in (71); Hence if $\widetilde{G}(\Delta) > 0$ for all $\Delta \in \mathbb{T}_n$, then $\widetilde{G}(\Delta) > 0$ for all $\Delta \in \mathbb{T}_n \cup \mathbb{W}_n$.

Note that for $\widehat{\Theta}_n \in \mathcal{S}^p_{++} \cap \mathcal{S}^p_E$, we have $\widehat{\Delta} = \widehat{\Theta}_n - \widetilde{\Theta}_0 \in U_n(\widetilde{\Theta}_0)$. By Proposition 21 and the fact that $\widetilde{G}(\widehat{\Delta}) \leq \widetilde{G}(\underline{0}) = 0$ on event $\mathcal{E}$, we have the following: on event $\mathcal{E}$, if $\widetilde{G}(\Delta) > 0, \forall \Delta \in \mathbb{T}_n$ then $\|\widehat{\Delta}\|_F < M r_n$, given that $\widehat{\Delta} \in U_n(\widetilde{\Theta}_0) \setminus (\mathbb{T}_n \cup \mathbb{W}_n)$. Therefore

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{\Delta}\|_F \geq M r_n\right) &\leq \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{E}) \cdot \mathbb{P}\left(\|\widehat{\Delta}\|_F \geq M r_n | \mathcal{E}\right) \\
&= \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{E}) \cdot (1 - \mathbb{P}\left(\|\widehat{\Delta}\|_F < M r_n | \mathcal{E}\right)) \\
&\leq \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{E}) \cdot (1 - \mathbb{P}\left(\widetilde{G}(\Delta) > 0, \forall \Delta \in \mathbb{T}_n | \mathcal{E}\right)) \\
&\leq \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{E}) \cdot (1 - \mathbb{P}(\mathcal{X}_0 | \mathcal{E})) \\
&= \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{X}_0^c \cap \mathcal{E}) \leq \mathbb{P}(\mathcal{E}^c) + \mathbb{P}(\mathcal{X}_0^c) \\
&\leq \frac{c}{p^2} + \frac{1}{\max\{p,n\}^2} \leq \frac{c+1}{p^2}.
\end{aligned}$$

We thus establish that the theorem holds. ■

D.3 Frobenius norm for the covariance matrix

We use the bound on $\left\|\widehat{\Theta}_n(E) - \Theta_0\right\|_F$ as developed in Theorem 14; in addition, we strengthen the bound on Mr_n in (77) in (80). Before we proceed, we note the following bound on bias of $(\widetilde{\Theta}_0)^{-1}$.

Remark 22 *Clearly we have on event $\mathcal{E}$, by* (75)

$$\left\|(\widetilde{\Theta}_0)^{-1} - \Sigma_0\right\|_F \;\le\; \frac{\|\Theta_{0,\mathcal{D}}\|_F}{\varphi_{\min}(\widetilde{\Theta}_0)\varphi_{\min}(\Theta_0)} \le \frac{32 C_{\text{bias}}\sqrt{2S_{0,n}\log p/n}}{31\underline{c}^2} \tag{79}$$

Proof of Theorem 15. Suppose event $\mathcal{E} \cap \mathcal{X}_0$ holds. Now suppose

$$n > \left(\frac{16}{7\underline{c}} \cdot \frac{9}{2\underline{k}^2}\right)^2 \left(C_3 + \frac{32}{31\underline{c}^2}\right)^2 \max\left\{(p + 2|E|)\log(n),\; C_{\text{bias}}^2 2S_{0,n}\log p\right\}$$

which clearly holds given (63). Then in addition to the bound in (77), on event $\mathcal{E} \cap \mathcal{X}_0$, we have

$$Mr_n < 7\underline{c}/16, \tag{80}$$

for r_n as in (71). Then, by Theorem 14, for the same M as therein, on event $\mathcal{E} \cap \mathcal{X}_0$, we have

$$\left\|\widehat{\Theta}_n(E) - \Theta_0\right\|_F \le (M+1)\max\left\{\sqrt{(p + 2|E|)\log(n)/n},\; C_{\text{bias}}\sqrt{2S_{0,n}\log(p)/n}\right\}$$

given that sample bound in (61) is clearly satisfied. We now proceed to bound $\left\|\widehat{\Sigma}_n - \Sigma_0\right\|_F$ given (62). First note that by (80), we have on event $\mathcal{E} \cap \mathcal{X}_0$ for $M > 7$

$$\begin{aligned}\varphi_{\min}(\widehat{\Theta}_n(E)) &\ge \varphi_{\min}(\Theta_0) - \left\|\widehat{\Theta}_n - \Theta_0\right\|_2 \ge \varphi_{\min}(\Theta_0) - \left\|\widehat{\Theta}_n - \Theta_0\right\|_F \\ &\ge \underline{c} - (M+1)r_n > \underline{c}/2.\end{aligned}$$

Now clearly on event $\mathcal{E} \cap \mathcal{X}_0$, (64) holds by (62) and

$$\left\|\widehat{\Sigma}_n(E) - \Sigma_0\right\|_F \;\le\; \frac{\left\|\widehat{\Theta}_n(E) - \Theta_0\right\|_F}{\varphi_{\min}(\widehat{\Theta}_n(E))\varphi_{\min}(\Theta_0)} < \frac{2}{\underline{c}^2}\left\|\widehat{\Theta}_n(E) - \Theta_0\right\|_F$$

■

D.4 Risk consistency

We now derive the bound on risk consistency. Before proving Theorem 16, we first state two lemmas given the following decomposition of our loss in terms of the risk as defined in (17):

$$0 \le R(\widehat{\Theta}_n(E)) - R(\Theta_0) = (R(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0)) + (R(\widetilde{\Theta}_0) - R(\Theta_0)) \tag{81}$$

where clearly $R(\widehat{\Theta}_n(E)) \ge R(\Theta_0)$ by definition. It is clear that $\widetilde{\Theta}_0 \in \mathcal{S}_n$ for $\mathcal{S}_n$ as defined in (58), and thus $\widehat{R}_n(\widetilde{\Theta}_0) \ge \widehat{R}_n(\widehat{\Theta}_n(E))$ by definition of $\widehat{\Theta}_n(E) \;=\; \arg\min_{\Theta \in \mathcal{S}_n} \widehat{R}_n(\Theta)$.

We now bound the two terms on the RHS of (81), where clearly $R(\widetilde{\Theta}_0) \ge R(\Theta_0)$.

Lemma 23 *On event $\mathcal{E}$, we have for $C_{\text{bias}}, \Theta_0, \widetilde{\Theta}_0$ as in Theorem 14,*

$$0 \leq R(\widetilde{\Theta}_0) - R(\Theta_0) \leq (32/(31\underline{c}))^2 C_{\text{bias}}^2 \frac{2S_{0,n}\log p}{2n} \leq (32/(31\underline{c}))^2 \cdot r_n^2/2 \leq M r_n^2/8$$

for r_n as in (71), where the last inequality holds given that $M \geq 9/2(C_3 + 32/(31\underline{c}^2))$.

Lemma 24 *Under $\mathcal{E} \cap \mathcal{X}_0$, we have for r_n as in (71) and M, C_3 as in Theorem 14*

$$R(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0) \leq M C_3 r_n^2.$$

Proof of Theorem 16. We have on $\mathcal{E} \cap \mathcal{X}_0$, for r_n is as in (71)

$$R(\widehat{\Theta}_n(E)) - R(\Theta_0) = (R(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0)) + (R(\widetilde{\Theta}_0) - R(\Theta_0)) \leq M r_n^2 (C_3 + 1/8)$$

as desired, using Lemma 23 and 24. ■

Proof of Lemma 23. For simplicity, we use Δ_0 as a shorthand for the rest of our proof:

$$\Delta_0 := \Theta_{0,\mathcal{D}} = \widetilde{\Theta}_0 - \Theta_0.$$

We use $\widetilde{B}$ as a shorthand for

$$\text{vec}\Delta_0{}^T \left(\int_0^1 (1-v)(\Theta_0 + v\Delta_0)^{-1} \otimes (\Theta_0 + v\Delta_0)^{-1} dv \right) \text{vec}\Delta_0,$$

where $\otimes$ is the Kronecker product. First, we have for $\widetilde{\Theta}_0, \Theta_0 \succ 0$

$$\begin{aligned} R(\widetilde{\Theta}_0) - R(\Theta_0) &= \text{tr}(\widetilde{\Theta}_0\Sigma_0) - \log|\widetilde{\Theta}_0| - \text{tr}(\Theta_0\Sigma_0) + \log|\Theta_0| \\ &= \text{tr}((\widetilde{\Theta}_0 - \Theta_0)\Sigma_0) - \left(\log|\widetilde{\Theta}_0| - \log|\Theta_0|\right) := \widetilde{B} \geq 0 \end{aligned}$$

where $\widetilde{B} = 0$ holds when $\|\Delta_0\|_F = 0$, and in the last equation, we bound the difference between two $\log|\cdot|$ terms using the Taylor's formula with integral remainder following that in proof of Theorem 14; Indeed, it is clear that on $\mathcal{E}$, we have

$$\Theta_0 + v\Delta_0 \succ 0 \ \text{ for } \ v \in (-1, 2) \supset [0, 1]$$

given that $\varphi_{\min}(\Theta_0) \geq \underline{c}$ and $\|\Delta_0\|_2 \leq \|\Delta_0\|_F \leq \underline{c}/32$ by (60). Thus $\log|\Theta_0 + v\Delta_0|$ is infinitely differentiable on the open interval $I \supset [0, 1]$ of v. Now, the Taylor expansion gives

$$\begin{aligned} \log|\Theta_0 + \Delta_0| - \log|\Theta_0| &= \frac{d}{dv}\log|\Theta_0 + v\Delta_0||_{v=0}\Delta_0 + \int_0^1 (1-v)\frac{d^2}{dv^2}\log|\Theta_0 + v\Delta_0| dv \\ &= \text{tr}(\Sigma_0\Delta_0) - \widetilde{B} \end{aligned}$$

where $\text{tr}(\Sigma_0\Delta_0) = \text{tr}((\widetilde{\Theta}_0 - \Theta_0)\Sigma_0)$ by symmetry. We now obtain an upper bound on $\widetilde{B} \geq 0$. Clearly, we have on event $\mathcal{E}$, Lemma 23 holds given that

$$\widetilde{B} \leq \|\Delta_0\|_F^2 \cdot \varphi_{\max}\left(\int_0^1 (1-v)(\Theta_0 + v\Delta_0)^{-1} \otimes (\Theta_0 + v\Delta_0)^{-1} dv \right)$$

where $\|\Delta_0\|_F^2 \leq C_{\text{bias}}^2 2S_{0,n} \log(p)/n$ and

$$\begin{aligned}
&\varphi_{\max}\left(\int_0^1 (1-v)(\Theta_0+v\Delta_0)^{-1} \otimes (\Theta_0+v\Delta_0)^{-1} dv\right)\\
&\leq \int_0^1 (1-v)\varphi^2_{\max}(\Theta_0+v\Delta_0)^{-1} dv \leq \sup_{v\in[0,1]} \varphi^2_{\max}(\Theta_0+v\Delta_0)^{-1} \int_0^1 (1-v)dv\\
&= \frac{1}{2}\sup_{v\in[0,1]} \frac{1}{\varphi^2_{\min}(\Theta_0+v\Delta_0)} = \frac{1}{2\inf_{v\in[0,1]} \varphi^2_{\min}(\Theta_0+v\Delta_0)}\\
&\leq \frac{1}{2\left(\varphi_{\min}(\Theta_0) - \|\Delta_0\|_2\right)^2} \leq \frac{1}{2\left(31\underline{c}/32\right)^2}
\end{aligned}$$

where clearly for all $v \in [0,1]$, we have $\varphi^2_{\min}(\Theta_0+v\Delta_0) \geq \left(\varphi_{\min}(\Theta_0) - \|\Delta_0\|_2\right)^2 \geq (31\underline{c}/32)^2$, given $\varphi_{\min}(\Theta_0) \geq \underline{c}$ and $\|\Delta_0\|_2 \leq \|\Theta_{0,\mathcal{D}}\|_F \leq \underline{c}/32$ by (60). ■

Proof of Lemma 24. Suppose $R(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0) < 0$, then we are done. Otherwise, assume $R(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0) \geq 0$ throughout the rest of the proof. Define

$$\widehat{\Delta} := \widehat{\Theta}_n(E) - \widetilde{\Theta}_0,$$

which by Theorem 14, we have on event $\mathcal{E} \cap \mathcal{X}_0$, and for M as defined therein,

$$\left\|\widehat{\Delta}\right\|_F := \left\|\widehat{\Theta}_n(E) - \widetilde{\Theta}_0\right\|_F \leq M r_n.$$

We have by definition $\widehat{R}_n(\widehat{\Theta}_n(E)) \leq \widehat{R}_n(\widetilde{\Theta}_0)$, and hence

$$\begin{aligned}
0 \leq R(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0) &= R(\widehat{\Theta}_n(E)) - \widehat{R}_n(\widehat{\Theta}_n(E)) + \widehat{R}_n(\widehat{\Theta}_n(E)) - R(\widetilde{\Theta}_0)\\
&\leq R(\widehat{\Theta}_n(E)) - \widehat{R}_n(\widehat{\Theta}_n(E)) + \widehat{R}_n(\widetilde{\Theta}_0) - R(\widetilde{\Theta}_0)\\
&= \text{tr}(\widehat{\Theta}_n(E)(\Sigma_0 - \widehat{S}_n)) - \text{tr}(\widetilde{\Theta}_0(\Sigma_0 - \widehat{S}_n))\\
&= \text{tr}((\widehat{\Theta}_n(E) - \widetilde{\Theta}_0)(\Sigma_0 - \widehat{S}_n)) = \text{tr}((\widehat{\Delta})(\Sigma_0 - \widehat{S}_n))
\end{aligned}$$

Now, conditioned on event $\mathcal{X}_0$, by Lemma 8

$$\max_{j,k} |\widehat{S}_{n,jk} - \sigma_{0,jk}| \leq C_3\sqrt{\log(n)/n} := \delta_n$$

and thus on $\mathcal{E} \cap \mathcal{X}_0$ (with probability at least $\mathbb{P}(\mathcal{E}) - \frac{1}{\max\{p,n\}^2}$), we have by Cauchy-Schwartz,

$$\begin{aligned}
\left|\text{tr}\left(\widehat{\Delta}(\widehat{S}_n - \Sigma_0)\right)\right| &\leq \delta_n \left|\widehat{\Delta}\right|_1 \leq \delta_n\sqrt{\left\|\widehat{\Delta}\right\|_0}\left\|\widehat{\Delta}\right\|_F \leq \delta_n\sqrt{p+2|E|}\left\|\widehat{\Delta}\right\|_F\\
&\leq M r_n C_3 \sqrt{\log(n)/n}\sqrt{p+2|E|} \leq M C_3 r_n^2
\end{aligned}$$

where $\left\|\widehat{\Delta}\right\|_0 \leq p + 2|E|$ by definition, and r_n is as defined in (71). ■

Appendix E. Oracle inequalities for the Lasso

In this section, we consider recovering $\beta \in \mathbb{R}^p$ in the following linear model:

$$Y = X\beta + \epsilon, \tag{82}$$

where X follows (1) and $\epsilon \sim N(0, \sigma^2 I_n)$. Recall given λ_n, the Lasso estimator for $\beta \in \mathbb{R}^p$ is defined as:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \tag{83}$$

which corresponds to the regression function in (10) by letting $Y := X_i$ and $X := X_{\cdot \setminus i}$ where $X_{\cdot \setminus i}$ denotes columns of X without i. Define

$$\sum_{i=1}^{p} \min(\beta_i^2, \lambda^2\sigma^2) \leq s_0 \lambda^2 \sigma^2, \text{ where } \lambda = \sqrt{2\log p/n}. \tag{84}$$

We now state Theorem 25, which may be of independent interests; here we derive a tighter bound for the Lasso estimator in terms of ℓ_2 convergence rate than that in Bickel et al. (2009) under slightly different RE conditions, see discussions below. Our bounds depend on the *actual* sparsity s_0 as defined in (84) rather than $s = |\operatorname{supp}(\beta)|$ as in Bickel et al. (2009) (cf. Theorem 7.2). A similar result has been shown in Zhou (2010b) for deterministic design matrices that satisfy the RE condition on X, where $\Lambda_{\max}(2s)$ as defined in (38) is assumed to be bounded. We now bound the correlation between the noise and covariates of X for $X \in \mathcal{X}$, where we also define a constant $\lambda_{\sigma,a,p}$ which is used throughout the rest of this paper. For $X \in \mathcal{F}(\theta)$ as defined in (34), let

$$\mathcal{T}_a := \left\{ \epsilon : \left\| \frac{X^T \epsilon}{n} \right\|_\infty \leq (1+\theta)\lambda_{\sigma,a,p}, \text{ where } X \in \mathcal{F}(\theta), \text{ for } 0 < \theta < 1 \right\}, \tag{85}$$

where $\lambda_{\sigma,a,p} = \sigma\sqrt{1+a}\sqrt{(2\log p)/n}$, where $a \geq 0$; we have (cf. Lemma 26)

$$\mathbb{P}(\mathcal{T}_a) \geq 1 - (\sqrt{\pi \log p}\, p^a)^{-1}; \tag{86}$$

In fact, for such a bound to hold, we only need $\frac{\|X_j\|_2}{\sqrt{n}} \leq 1 + \theta, \forall j$ to hold in $\mathcal{F}(\theta)$. The proof appears in Zhou (2010a).

Theorem 25 (Oracle inequalities of the Lasso) Zhou (2010a) *Let $Y = X\beta + \epsilon$, for ϵ being i.i.d. $N(0, \sigma^2)$ and let X follow (1). Let s_0 be as in (84) and T_0 denote locations of the s_0 largest coefficients of β in absolute values. Suppose that $RE(s_0, 4, \Sigma_0)$ holds with $K(s_0, 4, \Sigma_0)$ and $\rho_{\min}(s) > 0$. Fix some $1 > \theta > 0$. Let β_{init} be an optimal solution to (83) with*

$$\lambda_n = d_0 \lambda \sigma \geq 2(1+\theta)\lambda_{\sigma,a,p} \tag{87}$$

where $a \geq 1$ and $d_0 \geq 2(1+\theta)\sqrt{1+a}$. Let $h = \beta_{\text{init}} - \beta_{T_0}$. Define

$$\mathcal{X} := \mathcal{R}(\theta) \cap \mathcal{F}(\theta) \cap \mathcal{M}(\theta).$$

Suppose that n satisfies (41), and (42) for $m = s$. Then on $\mathcal{T}_a \cap \mathcal{X}$, we have

$$\begin{aligned}
\|\beta_{\text{init}} - \beta\|_2 &\le \lambda_n \sqrt{s_0}\sqrt{2D_0^2 + 2D_1^2 + 2} := \lambda\sigma\sqrt{s_0}\, d_0 \sqrt{2D_0^2 + 2D_1^2 + 2}, \\
\left\|h_{T_0^c}\right\|_1 &\le D_1 \lambda_n s_0 := D_1 d_0 \lambda \sigma s_0,
\end{aligned}$$

where D_0 and D_1 are defined in (88) to (90) respectively, and $\mathbb{P}\left(\mathcal{X} \cap \mathcal{T}_a\right) \ge 1 - 3\exp(-\bar{c}\theta^2 n/\alpha^4) - (\sqrt{\pi \log p}\, p^a)^{-1}$.

Let T_1 denote the s_0 largest positions of h in absolute values outside of T_0; Let $T_{01} := T_0 \cup T_1$. The proof of Theorem 25 yields the following bounds on $\mathcal{X} \cap \mathcal{T}_a$: $\|h_{T_{01}}\|_2 \le D_0 d_0 \lambda \sigma \sqrt{s_0}$ where

$$D_0 = \max\left\{\frac{D}{d_0},\ \sqrt{2}\left(2(1+\theta)\frac{K(s_0,4,\Sigma_0)\sqrt{\rho_{\max}(s-s_0)}}{(1-\theta)d_0} + \frac{3K^2(s_0,4,\Sigma_0)}{(1-\theta)^2}\right)\right\}, \quad (88)$$

$$\text{where } D = \frac{3(1+\theta)\sqrt{\rho_{\max}(s-s_0)}}{(1-\theta)\sqrt{\rho_{\min}(2s_0)}} + \frac{2(1+\theta)^4 \rho_{\max}(3s_0)\rho_{\max}(s-s_0)}{d_0(1-\theta)^2\rho_{\min}(2s_0)}, \quad \text{and} \quad (89)$$

$$D_1 = \max\left\{\frac{4(1+\theta)^2\rho_{\max}(s-s_0)}{d_0^2}, \left(\frac{(1+\theta)\sqrt{\rho_{\max}(s-s_0)}}{d_0} + \frac{3K(s_0,4,\Sigma_0)}{2(1-\theta)}\right)^2\right\} (90)$$

We note that implicit in these constants, we have used the concentration bounds for $\Lambda_{\max}(3s_0)$, $\Lambda_{\max}(s-s_0)$ and $\Lambda_{\min}(2s_0)$ as derived in Theorem 6, given that (40) holds for $m \le \max(s, (k_0+1)s_0)$, where we take $k_0 > 3$. In general, these maximum sparse eigenvalues as defined above will increase with s_0 and s; Taking this issue into consideration, we fix for $c_0 \ge 4\sqrt{2}$, $\lambda_n = d_0 \lambda \sigma$ where

$$d_0 \ge c_0(1+\theta)^2\sqrt{\rho_{\max}(s-s_0)\rho_{\max}(3s_0)} \ge 2(1+\theta)\sqrt{1+a},$$

where the second inequality holds for $a = 7$ as desired, given $\rho_{\max}(3s_0), \rho_{\max}(s-s_0) \ge 1$.

Thus we have for $\rho_{\max}(3s_0) \ge \rho_{\max}(2s_0) \ge \rho_{\min}(2s_0)$

$$\begin{aligned}
D/d_0 &\le \frac{3}{c_0(1+\theta)(1-\theta)\sqrt{\rho_{\max}(3s_0)}\sqrt{\rho_{\min}(2s_0)}} + \frac{2}{c_0^2(1-\theta)^2\rho_{\min}(2s_0)} \\
&\le \frac{3\sqrt{\rho_{\min}(2s_0)}}{c_0(1-\theta)^2\sqrt{\rho_{\max}(3s_0)}\rho_{\min}(2s_0)} + \frac{2}{c_0^2(1-\theta)^2\rho_{\min}(2s_0)} \\
&\le \frac{2(3c_0+2)K^2(s_0,4,\Sigma_0)}{c_0^2(1-\theta)^2} \le \frac{7\sqrt{2}K^2(s_0,4,\Sigma_0)}{8(1-\theta)^2}
\end{aligned}$$

which holds given that $\rho_{\max}(3s_0) \geq 1$, and $1 \leq \frac{1}{\sqrt{\rho_{\min}(2s_0)}} \leq \sqrt{2}K(s_0, k_0, \Sigma_0)$, and thus $\frac{1}{K^2(s_0,k_0,\Sigma_0)} \leq 2$ as shown in Lemma 27; Hence

$$\begin{aligned}
D_0 &\leq \max\left\{D/d_0, \frac{(4+3\sqrt{2}c_0)\sqrt{\rho_{\max}(s-s_0)\rho_{\max}(3s_0)}(1+\theta)^2 K^2(s_0,4,\Sigma_0)}{d_0(1-\theta)^2}\right\}, \\
&\leq \frac{7\sqrt{2}K^2(s_0,4,\Sigma_0)}{2(1-\theta)^2} \quad \text{and} \\
D_1 &= \max\left\{\frac{4(1+\theta)^2\rho_{\max}(s-s_0)}{d_0^2}, \left(\frac{(1+\theta)\sqrt{\rho_{\max}(s-s_0)}}{d_0} + \frac{3K(s_0,4,\Sigma_0)}{2(1-\theta)}\right)^2\right\} \\
&\leq \left(\frac{6}{4(1-\theta)} + \frac{1}{4}\right)^2 K^2(s_0,4,\Sigma_0) \leq \frac{49K^2(s_0,4,\Sigma_0)}{16(1-\theta)^2},
\end{aligned}$$

where for both D_1, we have used the fact that

$$\begin{aligned}
\frac{2(1+\theta)^2\rho_{\max}(s-s_0)}{d_0^2} &= \frac{2}{c_0^2(1+\theta)^2\rho_{\max}(3s_0)} \leq \frac{2}{c_0^2(1+\theta)^2\rho_{\min}(2s_0)} \\
&\leq \frac{4K^2(s_0,4,\Sigma_0)}{c_0^2(1+\theta)^2} \leq \frac{K^2(s_0,4,\Sigma_0)}{8}.
\end{aligned}$$

Appendix F. Misc bounds

Lemma 26 *For fixed design X with $\max_j \|X_j\|_2 \leq (1+\theta)\sqrt{n}$, where $0 < \theta < 1$, we have for $\mathcal{T}_a$ as defined in (85), where $a > 0$, $\mathbb{P}(\mathcal{T}_a^c) \leq (\sqrt{\pi \log p} p^a)^{-1}$.*

Proof Define random variables: $Y_j = \frac{1}{n}\sum_{i=1}^n \epsilon_i X_{i,j}$. Note that $\max_{1\leq j\leq p}|Y_j| = \|X^T\epsilon/n\|_\infty$. We have $\mathbb{E}(Y_j) = 0$ and $\textsf{Var}((Y_j)) = \|X_j\|_2^2\sigma^2/n^2 \leq (1+\theta)\sigma^2/n$. Let $c_1 = 1+\theta$. Obviously, Y_j has its tail probability dominated by that of $Z \sim N(0, \frac{c_1^2\sigma^2}{n})$:

$$\mathbb{P}(|Y_j| \geq t) \leq \mathbb{P}(|Z| \geq t) \leq \frac{2c_1\sigma}{\sqrt{2\pi n}t}\exp\left(\frac{-nt^2}{2c_1^2\sigma_\epsilon^2}\right).$$

We can now apply the union bound to obtain:

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq j\leq p}|Y_j| \geq t\right) &\leq p\frac{c_1\sigma}{\sqrt{n}t}\exp\left(\frac{-nt^2}{2c_1^2\sigma^2}\right) \\
&= \exp\left(-\left(\frac{nt^2}{2c_1^2\sigma^2} + \log\frac{t\sqrt{\pi n}}{\sqrt{2}c_1\sigma} - \log p\right)\right).
\end{aligned}$$

By choosing $t = c_1\sigma\sqrt{1+a}\sqrt{2\log p/n}$, the right-hand side is bounded by $(\sqrt{\pi \log p}p^a)^{-1}$ for $a \geq 0$. ■

Lemma 27 *(Zhou (2010a)) Suppose that $RE(s_0, k_0, \Sigma_0)$ holds for $k_0 > 0$, then for $m = (k_0{+}1)s_0$,*

$$\sqrt{\rho_{\min}(m)} \geq \frac{1}{\sqrt{2 + k_0^2} K(s_0, k_0, \Sigma_0)}; \quad \textit{and clearly}$$

$$\textit{if } \Sigma_{0,ii} = 1, \forall i, \quad \textit{then } 1 \geq \sqrt{\rho_{\min}(2s_0)} \geq \frac{1}{\sqrt{2} K(s_0, k_0, \Sigma_0)} \quad \textit{for } k_0 \geq 1.$$

References

BANERJEE, O., GHAOUI, L. E. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* **9** 485–516.

BICKEL, P. J. and LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, ”naive Bayes”, and some alternatives when there are many morevariables than observations. *Bernoulli* **10** 989–1010.

BICKEL, P. J. and LEVINA, E. (2008). Regulatized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732.

BÜHLMANN, P. and MEIER, L. (2008). Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36** 1534–1541.

CANDÈS, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

CHAUDHURI, S., DRTON, M. and RICHARDSON, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94** 1–18.

D'ASPREMONT, A., BANERJEE, O. and GHAOUI, L. E. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications* **30** 56–66.

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics* **3** 521–541.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**.

FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis* **98** 227–255.

HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse highdimensional regression. *Statistica Sinica* **18** 1603–1618.

HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.

JOHNSTONE, I. (2001). Chi-square oracle inequalities. *In State of the Art in Probability and Statistics, Festchrift for Willem R. van Zwet, M. de Gunst and C. Klaassen and A. van der Waart editors, IMS Lecture Notes - Monographs* **36** 399–418.

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *The Annals of Statistics* **37** 4254–4278.

LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *The Annals of Applied Statistics* **2** 245–263.

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis* **52** 374–393.

MEINSHAUSEN, N. (2008). A note on the Lasso for gaussian graphical model selection. *Statistics and Probability Letters* **78** 880–884.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34** 1436–1462.

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37** 246–270.

RAVIKUMAR, P., WAINWRIGHT, M., RASKUTTI, G. and YU, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. In *Advances in Neural Information Processing Systems*. MIT Press. Longer version in arXiv:0811.3628v1.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

RÜTIMANN, P. and BÜHLMANN, P. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic Journal of Statistics* **3** 1133–1160.

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2010). The adaptive and the thresholded Lasso for potentially misspecified models. ArXiv:1001.5176v3.

WEST, M., BLANCHETTE, C., DRESSMAN, H., HUANG, E., ISHIDA, S., SPANG, R., ZUZAN, H., JR., J. O., MARKS, J. and NEVINS, J. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS* **98** 11462–11467.

WILLE, A., ZIMMERMANN, P., VRANOVA, E., FÜRHOLZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIC, A., VON ROHR, P., THIELE, L., ZITZLER, E., GRUISSEM, W. and BÜHLMANN, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology* **5** R92.

WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94** 19–35.

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7** 2541–2563.

ZHOU, S. (2009). Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems 22*. MIT Press.

ZHOU, S. (2010a). Restricted eigenvalue conditions on subgaussian random matrices. Manuscript, earlier version in arXiv:0904.4723v2.

ZHOU, S. (2010b). Thresholded Lasso for high dimensional variable selection and statistical estimation. ArXiv:1002.1583v2, also available as University of Michigan, Department of Statistics Technical Report 511.

ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2008). Time varying undirected graphs. In *Proceedings of the 21st Annual Conference on Computational Learning Theory (COLT'08)*.

ZHOU, S., VAN DE GEER, S. and BÜHLMANN, P. (2009). Adaptive Lasso for high dimensional regression and gaussian graphical modeling. ArXiv:0903.2515.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429.

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36** 1509–1533.